Dissertation

The Intellectual Salmon Run:
Knowledge Transfer and Dynamics between Academia and Industry

Thomas Gurney

Rathenau Instituut

The **Rathenau Institute** promotes the formation of political and public opinion on science and technology. To this end, the Institute studies the organization and development of science systems, publishes about social impact of new technologies, and organizes debates on issues and dilemmas in science and technology.

The Intellectual Salmon Run:

Knowledge Transfer and Dynamics between Academia and Industry

Rathenau Instituut
Anna van Saksenlaan 51

Postal address:
Postbus 95366
2509 CJ Den Haag

Telefoon: +31 (0)70 342 15 42
Telefax: +31 (0)70 363 34 88
E-mail: info@rathenau.nl
Website: www.rathenau.nl

Publisher: Rathenau Instituut
Lay-out: Boven de Bank, Amsterdam
Coverphoto: H.H.
Printing: Drukkerij Quantes, Rijswijk

This book is printed on FSC certified paper.

ISBN/EAN: 978-90-77364-50-5

Preferred citation:
Gurney, T., The Intellectual Salmon Run: Knowledge Transfer and Dynamics between Academia and Industry, Den Haag, Rathenau Instituut 2013

This dissertation is based on research the author has done as an employee of the Rathenau Instituut. This text is the author's PhD thesis. The views in this dissertation are not necessarily the same as those of the Rathenau Instituut.

VRIJE UNIVERSITEIT

The Intellectual Salmon Run:

Knowledge Transfer and Dynamics between Academia and Industry

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. F.A. van der Duyn Schouten,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Sociale Wetenschappen
op dinsdag 21 januari 2014 om 11.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door

Thomas Michael Gurney

geboren te Harare, Zimbabwe

promotor: prof. dr. P.A.A. van den Besselaar

Thesis committee:
dr. Peter van der Sijde (VU, FSW, Afdeling ORG)
prof.dr. Tom Elfring (VU, FEWEB)
prof.dr. Marleen Huijsman (VU, FEWEB)
prof.dr. Robert Tijssen (U Leiden)
prof.dr. Bart van Looy (U Leuven)

Content

1 Introduction

1.1 Preamble

"The colour of the object illuminated partakes of the colour of that which illuminates it." Although likely mistranslated from the original text, it remains clear what Leonardo Da Vinci was trying to impart on students of art.

This phrase also applies to the world of technology and knowledge, and their transference. We understand the technologies currently in use or in development as the result of knowledge accumulated over time and applied in varied, and sometimes new, forms. Education and practice allow scientists and researchers to understand the phenomena they observe at a fundamental level, and to devise novel methods to apply their understanding of them. To paraphrase the previous quote, the light of knowledge and experience can craft an idea into something entirely new.

However, the knowledge that is generated in one locale frequently needs to be translated, transferred or transliterated to find meaningful application in another. In other words, in the dynamics of science and technology, the spawning grounds of theory and the hatching grounds of application are divided by an ocean of experience and time - and it is across this ocean we aim to swim.

1.2 Primary research question

The transfer of knowledge across the metaphorical ocean of experience and time is not radically different from the reality. The end-results of the vast interplay between individuals, firms, universities and environments - be they products, processes or ideas - follow convoluted paths, operating via metaphorical diffusion gradients. Unlike diffusion gradients seen in nature, the diffusion process in knowledge transfer does not spontaneously occur. It takes concerted effort, be it at the fine-grained level of two individuals communicating, or at the supra-national policy level. There remains uncertainty in the research that has been produced on knowledge transfer. Many studies have argued that knowledge transfer does occur, and this is stated as a matter of fact. However, there are still many questions surrounding the operationalisation of knowledge transfer. We still do not know what knowledge is transferred, from where and to whom, and how the transfer and reception work exactly. Adding to this, we do not know the conditions surrounding knowledge transfer. To clarify this is not only of scholarly interest, but also of interest to society in terms of innovation and innovation policy, higher education and science policy. Industry has a vested interest in this, as knowledge transfer between academia and industry provides a significant portion of the inspiration and knowledge they require to produce and develop products and services.

To deduce the processes and mechanisms involved in knowledge transfer, it is necessary to observe the specific quanta and carriers of knowledge. However, unlike a Heisenbergian interpretation, it is possible to analyse both the quanta and the carrier at the same time. To do this, it is necessary to define the three primary aspects of knowledge transfer. The first involves the knowledge itself - how was it generated, how has it developed and how is it primed for transfer. The second involves the 'sender' and 'receiver' of the information or knowledge - who are they and how has each contributed to the knowledge. And the third involves the environment - how have the conditions surrounding the knowledge facilitated a productive transfer.

The receptivity and application of knowledge, is known as its absorptive capacity, and is a major facet in knowledge transfer. Two key aspects of absorptive capacity contribute to the understanding of knowledge transfer: the capacities of the individual researcher and those of the firm and its environment. The structure and patterns of communication between a researcher and his or her environment is a key factor in describing the absorptive capacity of the individuals and firms involved. Communication between actors occurs in a specific environment, with specific environmental and social factors influencing the transfer of knowledge. In addition, complementing available knowledge assets by linking to and incorporating external knowledge, can spawn entirely new lines of thinking.

As such, the primary question I want to answer in this thesis is:
What knowledge elements are transferred from academia to industry, how are they transferred, and what factors influence this transfer?

The results should provide a base for future studies of the effectiveness of certain knowledge transfer strategies and their effect on firm development. In order to clarify the factors that influence knowledge transfer, the role of Science Parks is closely scrutinised, highlighting many of the positive aspects cited by proponents of such research infrastructures, but also bringing forward the mythical aspects.

1.3 Theoretical framework

1.3.1 Knowledge transfer

The concept of knowledge transfer between academia and industry at a theoretical level was thought to follow a linear process (Gibbons et al., 1994) and was held to be the standard for many years. There is a clear theoretical move towards a complex interplay between knowledge producers, users and environment including feedback paths, in 'systems of innovation' (Edquist & Hommen, 1999), including the triple helix notion of overlapping academic, industrial and governmental knowledge creation and stimulation (Etzkowitz, 1998; Etzkowitz & Leydesdorff, 2000). These systems of innovation theories are grand in design, covering geographic or technological areas (Hekkert et al., 2007) and address the meta-environment of science and technology interactions (Meyer, 2002). They function well to describe the optimal environmental conditions for effective policy in facilitating science and technology interactions.

In addition to these grand designs, the Bayh-Dole act in the USA and the EU Framework Programmes have both been credited with contributing to more productive interactions between science and technology (Caloghirou et al., 2001; Mowery et al., 2001). These policy and funding instruments were primarily aimed at the interaction space between industry and universities, and how universities could better orientate themselves (and the human capital they represent) towards application-driven science. Funding instruments available to European universities have become more project-orientated (Bonaccorsi, 2007; Lepori et al., 2007a; Lepori et al., 2007b) with greater industrial financial input (Gulbrandsen & Smeby, 2005). Universities have tended towards specialisation, especially on realisation of the need to capitalise on research conducted within the university, undergoing a "second academic revolution" where economic development is becoming embedded in the charter of a university (Etzkowitz, 1998; Etzkowitz et al., 2000). The

development of Technology Transfer Offices (TTOs) to facilitate the capitalisation of research (Markman et al., 2005a; Markman et al., 2005b), and the introduction of university practices aimed at facilitating the development of spin-offs (Feldman et al., 2002) have become commonplace amongst universities.

Proponents of systems of innovation theories would say that the specific knowledge transfer processes occur against the background of a national system of innovation. For analytical purposes, however, the scale of these systems becomes a hindrance. They hinder examining specific developments and origins of technologies and the associated science. In other words, the system of innovation serves to condition the knowledge transfer processes at the micro-level rather than describing the knowledge transfer processes that occur.

On a practical level, knowledge transfer and associated mechanisms typically focus on mediums. Examples of which include technology or skills (Steffensen et al., 2000) where participants receive the knowledge required to perform tasks with a certain technology through the construction and utilisation of that technology itself. Another example includes contracts or collaborations (Agrawal et al., 2006) in which participants working in close proximity (be it cognitive or physical) supplement their current knowledge and skills with those of their collaborators or contractors/ contractees. Transfer mediums are typically codified in publications and patents or can be tacit (Cohen et al., 2002). Commonly used indicators are based on patent and publication data, as publications are considered to be the most visible outcome of scientific research and patent applications provide detailed evidence of technological progress (Tijssen, 2002).

Knowledge transfer has typically been addressed in the extant literature as something that occurs as matter of fact. Breschi and Lissoni (2001) even argue against the 'stylised fact' of localised knowledge spillovers, in which it is assumed that knowledge spillovers occur. There are more complex processes at work within knowledge transfer, other than merely assuming or expecting occurrence. To start, the actual knowledge elements transferred serve as a black box and what is missing is an adequate methodology for *quantifying* the tracks and knowledge being transferred. In this thesis, we aim to provide some understanding and clarification in this direction, leading to the first sub-question relating to the quantification of knowledge transfer, and the actors involved. Specifically, *how can we identify knowledge elements and their attributes in an operational way, and what elements are transferred between actors?*

To do this, it is necessary to understand first the ability to transfer and receive knowledge, and this is highly dependent on the infrastructure of the supporting knowledge platforms. Key to these knowledge platform infrastructures is the receptivity or absorptive capacity (Cohen & Levinthal, 1990) of the recipient.

1.3.2 Absorptive capacity and academic spin-offs

Absorptive capacity, or the ability to recognise the utility of new information and to translate and apply it commercially (Cohen & Levinthal, 1990) best describes the sender and receiver aspect of knowledge transfer. For both the individual and firm, the knowledge assets (Nonaka, 1994) in place influence the ability to recognise the utility of new knowledge, as well as the ability to merge new knowledge with current knowledge to produce novel artefacts, processes and

understanding. Absorptive capacity may be considered both in terms of the individuals comprising the firm, and as the firm itself. As stated by Cohen and Levinthal (1990), "Beyond diverse knowledge structures, the sort of knowledge that individuals should possess to enhance organizational absorptive capacity is also important. Critical knowledge does not simply include substantive, technical knowledge; it also includes awareness of where useful complementary expertise resides within and outside the organization" (p.133). In this manner a key aspect is the communication between the firm and the outside world. To this end, key individuals should be considered, namely the star (Zucker & Darby, 1996), core (Furukawa & Goto, 2006), or Pasteur scientist (Stokes, 1997) as they are best positioned to recognise the current knowledge platforms of the firm and how best to supplement them (Baba et al., 2009). Additionally, market demands may affect the search for new knowledge or ideas to incorporate (Langrish et al., 1972), including the best way to integrate them into the current knowledge platform, such as through acquisition of staff or new in-house R&D efforts.

The concept of absorptive capacity was expanded on significantly by Zahra & George (2002) who distinguished between potential and realised absorptive capacity. This is done by adding four operational dimensions to the definitions of absorptive capacity. These include, for potential absorptive capacity, *acquisition* - which necessitates the taking of stock or inventory of the current assets and knowledge platforms; and *assimilation* - which requires the knowledge intended to be brought in not only to be understood theoretically but also in terms of its place within current knowledge platforms. In realised absorptive capacity, the dimensions of *transformation* - which includes the ability to create novel knowledge by adding external knowledge to the current platform, and *exploitation* - in which results of the combined aforementioned dimensions are brought to light. Examples of exploitation could include, but are not limited to, patent applications, scientific publications or new work processes.

To address some of the complex processes in measuring knowledge transfer and absorptive capacity, studies frequently involve academic spin-offs because they provide the clearest identifiable path of knowledge transfer, where an idea can be followed from its inception to its commercial roll-out through a specific individual or group. Spin-offs embody an idea which was developed in academia and deemed to be commercially viable, but they require a dedicated entity to manifest. Studies involving spin-offs generally focus on the typologies of the firms and progenitor universities (Jones-Evans, 1995; Mustar et al., 2006; Westhead & Storey, 1995) or the environments they settle in - most commonly a choice between on or off a Science Park (Dettwiler et al., 2006; Felsenstein, 1994; Fukugawa, 2006; Hansson et al., 2005). Overall, studies such as the above provide indications of the roles of the individuals involved with the knowledge transfer, as well as the source and end-user environments of the knowledge, but do not examine their effects on the actual knowledge elements being transferred.

For studies relying on scientific output, their analyses must be based on accurate data, especially when using scientific publications. In researching the precise knowledge elements being transferred, the scientific publications of the person(s) under study must be positively identified as belonging to that individual and not another researcher of the same name. This is a common problem in studies that utilise scientific publications. With the rise of the Asian science systems, and the associated low variance in Asian researcher names, this problem is likely to get worse. To tackle

this, we strongly require an understanding of the problems related to name ambiguity, plus a reliable and effective process to accurately disambiguate the sometimes vast number of publications.

1.3.3 Disambiguation

An automated approach to disambiguation is necessary, which is particularly important now that the scale and scope of databases is increasing dramatically (Cassiman et al., 2007; Moed et al., 2004). Automated methods tend to follow either a computer science or a sociological/linguistic approach or a combination of the two. Essentially, similarities between two authors and their outputs are calculated using various models. These include Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) (Magerman et al., 2010; Song et al., 2007). These approaches have been successful to a degree but most suffer from a common drawback, that of data discarding. When comparing output records between two entities, it is common to discard a record that does not possess the necessary field for which the comparison was conducted. Therefore, the source databases used for many of these studies, such as Thompson Reuters ISI Web of Science, are not perfect in themselves because data may already be missing. For example, studies utilising key words suffer if any records are missing their keywords. Another example is that of using co-author similarity to determine if two records are from the same author. When using co-authors, how to handle records with only one author i.e. no co-authors? In practice, these records are discarded, to the detriment of the resulting precision and recall of the algorithm. For large, comparative studies the source data needs to be taken into consideration, and disambiguation is a necessity.

To address the issue of data accuracy, the second sub-question of this thesis is: ***How can we disambiguate researchers with an effective balance between precision and recall?***

Once we are able to construct databases in which the accuracy of the publications contained within are to our satisfaction, we can analyse the knowledge elements being transferred with a higher degree of confidence. Returning to the transfer of knowledge elements, in order to link absorptive capacity and spin-offs, we examine a common route to enabling the infrastructure for absorptive capacity. This lies in the choice of location for an academic spin-off (Volberda et al., 2010). For spin-offs, the environment is crucial for absorptive capacity to occur. Cohen and Levinthal (1990) state that absorptive capacity is the "ability to identify, assimilate, and exploit knowledge from the environment". The environment offers firms a *choice* of knowledge, and *access* to an environment is often the first step for firms stepping outside the university. This is in line with resource-based theory, in which academic spin-offs require access to different resources including capital, personnel, space and knowledge (Dettwiler et al., 2006; Klofsten, 1999; Löfsten & Lindelöf, 2005). For academic spin-offs, an environment that provides this is often a Science Park.

1.3.4 Science Parks

There is a substantial body of literature that describes Science Parks as providing an environment to promote knowledge transfer and interactions between firms, universities and small labs (Das & Teng, 1997; Löfsten & Lindelöf, 2005; Siegel et al., 2003); they provide a contact space between the 'fast applied science' of industry and the 'slow basic science' of the university (Quintas et al., 1992), and provide a technological platform for economic development at a regional or national level (Castells & Hall, 1994; Felsenstein, 1994; Phillimore, 1999).

Each Science Park may have specific origins but there are three general growth mechanisms involved. These include: government-directed mechanisms; agglomerative effects; and new firm creation and self-renewal (Koh et al., 2005). The agglomerative effects and new firm creation and self-renewal mechanisms are strongly linked to ease of access to qualified personnel, primarily through universities located nearby. Access to this sort of personnel is a boon for firm-founders looking to transfer research conducted in academia to application or commercialisation in industry.

Science park locations primarily appeal to firms which are either industry-based spin-outs, or academic spin-offs. There are three distinct reasons at the heart of the motivations of each type of firm to join a Science Park (Westhead & Batstone, 1998). The first of which is related to neoclassical theory in which transport, labour costs, distance to customers, and agglomeration economies are influential. The second set of reasons are related to behavioural aspects including the presence of mediators, gatekeepers or information channels in the form of the Science Park management. Additionally, the reputational advantages of situating in a Science Park play a large role in influencing firm founders to locate in a park. Most importantly for this thesis, the third set of reasons relate to structuralist approaches, including access to an innovative, networked environment, in which the presence of a Higher Education Institution plays a central role.

However, in all the literature on Science Parks, the idiosyncrasies of each Science Park add complexity. The convoluted histories and serendipitous moments (Dodgson & Hinze, 2000) of firms located on Science Parks adds force difficult questions to arise in studies comparing Science Parks. As such, considering academic spin-offs most frequently choose to locate in a Science Park, we examine the environment and associated networks found in a Science Park that facilitate or affect knowledge transfer, rather than comparing Science Parks. From this, the third and last sub-question: *What resources, and from which actors and operational spheres, contribute most significantly to the development of an academic spin-off and its host technology?*

1.4 Sub-Questions

Given the theoretical bases touched upon in the above sections and the various methodological issues with measuring knowledge transfer, the general question formulated in the introduction can be split into sub-questions that will be addressed in the chapters to follow. This thesis aims to provide a tool kit of methods and concepts to help answer each of the sub-questions and, ultimately, the primary research question. To recap, the sub-questions are:

- How can we identify knowledge elements and their attributes in an operational way, and what elements are transferred between actors?
- What resources, and from which actors and operational spheres, contribute most significantly to the development of an academic spin-off and its host technology?

Related to the accuracy of the data for the above-mentioned sub-questions:

- How can we disambiguate researchers with an effective balance between precision and recall?

1.5 Selected case studies

The selection of case studies for answering the primary and sub-questions was based on primarily isolating, or controlling for, extraneous circumstances. Beginning with the choice of academic rather than industrial spin-offs, academic spin-offs represent the immediate transfer of ideas generated in academia to industry, whereas industry spin-offs do not have any immediate links with academic research. The incentives against failure also differ between the two as the financial support given to academic spin-offs as provided by the university is minimal as compared to the relatively larger financial backing of an industrial parent. This affects the level of responsibility felt by the individual, as the success or failure of the firm lies solely on the firm founder's shoulders, in both a personal and academic sense. For industrial spin-offs, the responsibility is still considerable, though to a lesser extent. The motivation to succeed in an academic spin-off is also driven by the individual, whereas with an industrial spin-off the parent firm is arguably more reliant on its success.

To investigate the environment surrounding the transfer of knowledge from academia to industry, the Science Park is often considered the best option for knowledge-intensive spin-offs to locate, at least initially. The presence of a university near to a Science Park is promoted for firm founders, as is the ready access to expertise that it represents should the need arise. From the perspective of the university, a Science Park may act as an important external conduit to valorisation of research conducted within the university. A Science Park arguably represents such an environment, as it provides an extension of the university's research capabilities by providing infrastructure and organisational aspects which it would not otherwise have space or budget for.

This thesis will refer to a specific individual, a collection of firms and a specific Science Park which have been selected to demonstrate the effectiveness of the tool kit of methods and concepts developed. The individual scientist used to answer sub-question *How can we identify knowledge elements and their attributes in an operational way, and what elements are transferred between actors?* is a prolific academic scientist and the founder of a successful biotechnology-orientated firm. Professor Nakamura, of the University of Tokyo and founder of Oncotherapy Sciences Ltd., provides a perfect case study to test our methodology in linking scientific publications and patent applications. With over 900 publications published through both the university and the firm, and over 100 patents granted with either the university or the firm as assignee, he is an excellent test bed to examine the transfer of knowledge from academia to industry.

To answer the sub-question *What resources, and from which actors and operational spheres, contribute most significantly to the development of an academic spin-off and its host technology?* the firms we selected are drawn from Leiden Bioscience Park, located in Leiden, The Netherlands. This Science Park is biotechnology-orientated with a wide variety of service, research and product development companies. There are a few large multinational companies located within the grounds and its close physical proximity to Leiden University (and its teaching hospital, Leiden University Medical Centre) are attractants to firms looking to locate there. There have been significant investments in the infrastructure of the park and there is an active park management team, responsible for recruiting new firms to its premises. The firms from Leiden Bioscience Park were selected on three primary criteria: the firm was formed within the last 10 years; the firm was founded by a university or knowledge institute researcher; and lastly, the firm is in the field of life sciences and health. Following these criteria, we eventually were able to include nine firms in this

thesis. These criteria were deployed so as to ensure certain commonalities i.e. economic climate, scientific field, approximate qualifications of the firm founder and formation origins (specifically academic rather than corporate spin-offs).

To answer the sub-question *How can we disambiguate researchers with an effective balance between precision and recall?* we utilised a data set of publications and authors that had been pre-cleaned by hand. This data set was prepared by a project team within the PRIME ERA Dynamics project. The data set is a collection of 4979 articles, letters, notes and reviews featuring 5616 authors, within the field of heterogeneous catalysis. It was important to begin with a 'gold-standard' data set, in that it allowed us to test our algorithms with confidence, knowing that the results were based on clean data.

1.6 Methods used

The methods used in this thesis address the different requirements of the sub-questions and ultimately the primary research question.

To gain an understanding of the data, the data was first cleaned and disambiguated (both algorithmically and by hand). The method of algorithmic disambiguation outlined in Gurney et al., (2012) utilises the similarities between publications of the metadata as well as the level of contribution of individual authors to each publication. In this method, we incorporated a logistic regression and cluster detection process. Furthermore, to counter-act the common problem of data discarding in disambiguation processes, our approach incorporates all the available meta-data for each record. If specific metadata fields were blank or missing, the next optimum combi-nations of available metadata were used. This resulted in a very effective dynamic approach for comparing records, where the optimum available metadata were used. In some instances, records were compared to each other on completely different metadata, with each combination of metadata providing differing predictive abilities.

The similarities between publications and patent applications were based on three methods, the first developed by Van den Besselaar & Heimeriks (2006). In this method, rather than using only the title words or only the cited references to create a similarity matrix between publications, a combination of both was used to encompass both the cognitive foreground (title words) and the scientific background (cited references). This combination provided a clearer view as to the context and content of the publications in one similarity metric. The third method is based on the community detection algorithm of Blondel et al. (2008), the 'Louvain Method' as is it is commonly referred to. This clustering algorithm allowed us to quickly allocate clusters within the similarity matrices created using the combination of title word and cited references (the first method). The algorithm is what is known as a 'greedy' algorithm where the modularity of the network is iteratively built and tested to its optimum, resulting in each publication being assigned to discrete clusters. The third method in this combination is based on the visualisation techniques developed in Horlings & Gurney (2012), in which the academic life cycle of a researcher was mapped according to the similarities calculated by Van den Besselaar & Heimeriks' methodology and Blondel's cluster assignments and displayed as latitudinal and longitudinal coordinates on a equirectangular map, separated by cluster and ordered by year. This method of visualisation afforded us a unique view of the development of individual scientists and the collective publication output of firms.

An extension of this was added to include the patent applications of the firm or individual to the map whereby the non-patent literature references cited by the patent applications are grouped together with the publications in terms of their overall similarity.

To visualise the interactions between firm founder and other entities in the context of a Science Park, we build upon a method we developed in Lanciano-Morandat (2009), in which the context and nature of the interaction, along with the proximity to the Science Park are mapped in a circular fashion, with the Science Park forming the centre of the circle. This method allows for a visually simple approach that is data-rich at the same time.

1.7 Introduction to chapters

This thesis is structured by the progression of methodological approaches necessary to fully examine the transfer of knowledge from academia to industry, how it is transferred and what factors influence the transfer. Below are short descriptions of each of the chapters including their overall contributions to the research questions.

In Chapter 2, part of ensuring the accuracy of the data used in this thesis, was the requirement to disambiguate the output of the individuals under study. The use, and any meaningful results, of any scientometric or bibliometric approach require an initial data set to be as error-free as possible. The approach taken in this chapter does not stem from a computer science direction, involving hugely complex algorithms, but rather from a sociological perspective. The general approach taken to disambiguation was unique in that we both prevented data from being discarded because of missing data fields, but also used time and author-contribution differences to increase the predictive ability of the algorithm. The time difference between records greatly influences the degree of similarity between details such as title words or co-authors. By allowing for a probable decrease in similarity over time, the approach can utilise other metadata such as host journal or cited references to better effect. Varying contributions from authors in metadata selection plays a large part in how similar records are to one another. Allowing for these behavioural aspects of publishing makes the algorithms more capable of distinguishing between individuals. In addition to the data cleaning and processing utility of the disambiguation algorithms developed in this chapter, the sociological and behavioural aspects discussed lead to one of the more important insights into the mechanisms of knowledge transfer i.e. the collaborative aspect of knowledge production.

Chapters 3 and 4 provide a detailed view on the construction, identification and tracking of knowledge elements, with Chapter 3 providing the methodological approach to be used in Chapter 4. Chapter 3 introduces the ideas of concept clusters in order to aid with the delineation of endogenous versus exogenously sourced knowledge in the analysis of the transfer of knowledge elements. Concept clusters are a result of clustering techniques where the non-patent literature references (NPLRs) in an inventor/author's patent applications are grouped with the publications of that same inventor/author. This results in, at a fine-grained level, the identification of specific theories and methods utilised and cited by the patent applications in terms of clusters of publications. These clusters are composed of (i) publications authored by the inventor/author, and cited by the patent applications; (ii) publications authored by the inventor/author, and not cited by the patent applications; (iii) contain NPLRs not authored by the inventor/author; or (iv) contain a

heterogeneous mix of the three types. The proportional mix of these concept clusters allows us to infer the direct or indirect contributions of the inventor/author, either in terms of skill sets or knowledge, and at which stage in the inventor/author's career they were utilised or built upon.

This approach resulted in direct and indirect cognitive links between the technologies manifest in the patent applications and the scientific output, and thus capabilities, of the inventor/author. By highlighting the links between the patent applications and the underlying topics of research, it was possible to trace the development of an idea generated in academia, through transformation by the individual and eventual application in industry, as manifest by the patents and their embodied technologies. This methodology is applied to a case study in Chapter 4. This chapter uses the concept clusters approach to identify and analyse the role of the firm founder (Nakamura) and his co-inventors in facilitating the transfer of knowledge from academia to industry. Following the theoretical extensions of absorptive capacity by Zahra & George (2002), this chapter introduces a framework of descriptors to provide an understanding of the level of exogenously and endogenously generated knowledge required for the technologies of the firm and the role of the Nakamura as contributor to these technologies. The framework addresses the following: the reputational and applicability aspects of Nakamura's scientific work; the overall research trajectories of Nakamura in relation to the technologies; the degree of utilisation by the technologies of scientific fields outside Nakamura's expertise; the degree of shared knowledge features between Nakamura's research and the technologies; the level of input of Nakamura's co-inventors and co-authors; and finally the degree of knowledge incorporated and applied to Nakamura to the specific technologies.

In the penultimate chapter, Chapter 5, the quantitative approaches developed and expanded on in Chapter 3 and Chapter 4 are combined with a qualitative approach that we developed and deployed in an earlier study (Lanciano-Morandat et al., 2009). This qualitative approach was adapted to academic start-ups based at Leiden Bioscience Park in The Netherlands. Of equal importance to the social networks reported by the firms was their scientific development (analysed in a way as detailed in Chapters 3 and 4). The two models were combined to assess the role of the Science Park as an entity, and as an environment for successful firm development. A firm selection profile was built using the methodologies detailed in Chapter 3, based on the interactions reported by the firms' founders and the scientific and technical efforts (in terms of publishing and patenting) of the firm founders and their collaborating partners. The overall aim of this chapter is to examine the cognitive routes and developments of an idea generated in academia and exploited in industry, in relation to support by the Science Park. These aims are achieved by analysing the links between the firms' founders and their technological output including the scientific and technological links to local, regional and international HEIs and public research institutes. It closely examines the role of the firm founders' knowledge stocks, including the activity characteristics of their collaborators. The facilitating role of the Science Park administration is examined as well as the role of other firms located at the Science Park.

The sixth and final chapter summarises the findings of previous chapters and elucidates theoretical links between them in order to answer the primary research question. Additional considerations are made to explore future research possibilities and the implications for policy makers, industrial actors, universities and the founders of start-ups located in Science Parks.

1.8 References

Agrawal, A. et al., (2006). Gone But Not Forgotten: Labor Flows, Knowledge Spillovers, and Enduring Social Capital. *Journal of Economic Geography, 6*, 20.

Baba, Y. et al., (2009). How do collaborations with universities affect firms' innovative performance? The role of "Pasteur scientists" in the advanced materials field. *Research Policy, 38*(5), 756-764.

Blondel, V. D. et al., (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*, P10008.

Bonaccorsi, A. (2007). *Universities and strategic knowledge creation: specialization and performance in Europe*: Edward Elgar Publishing.

Breschi, S. & Lissoni, F. (2001). Knowledge Spillovers and Local Innovation Systems: A Critical Survey (Vol. 10, pp. 975-1005): Oxford Univ Press.

Caloghirou, Y. et al., (2001). University-industry cooperation in the context of the European framework programmes. *The Journal of Technology Transfer, 26*(1), 153-161.

Cassiman, B. et al., (2007). Measuring industry-science links through inventor-author relations: A profiling methodology. *Scientometrics, 70*(2), 379-391.

Castells, M. & Hall, P. (1994). *Technopoles of the World: The Making of 21st Century Industrial Complexes*: Routledge.

Cohen, W.M. et al., (2002). R&D spillovers, patents and the incentives to innovate in Japan and the United States. *Research Policy 31*, 1349-1367.

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

Das, T.K. & Teng, B.S. (1997). Time and Entrepreneurial Risk Behavior. *Entrepreneurship Theory and Practice, 22*(2), 69-71.

Dettwiler, P. et al., (2006). Utility of location: A comparative survey between small new technology-based firms located on and off Science Parks--Implications for facilities management. *Technovation, 26*(4), 506-517.

Dodgson, M. & Hinze, S. (2000). Indicators used to measure the innovation process: defects and possible remedies. *Research Evaluation, 9*(2), 101-114.

Edquist, C. & Hommen, L. (1999). Systems of innovation: theory and policy for the demand side. *Technology in Society, 21*(1), 63-79.

Etzkowitz, H. (1998). The norms of entrepreneurial science: cognitive effects of the new university-industry linkages. *Research Policy, 27*(8), 823-833.

Etzkowitz, H. & Leydesdorff, L. (2000). The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university-industry-government relations. *Research Policy, 29*(2), 109-123.

Etzkowitz, H. et al., (2000). The future of the university and the university of the future: evolution of ivory tower to entrepreneurial paradigm. *Research Policy, 29*(2), 313-330.

Feldman, M. et al., (2002). Equity and the technology transfer strategies of American research universities. *Management Science, 48*(1), 105-121.

Felsenstein, D. (1994). University-related Science Parks -- 'seedbeds' or 'enclaves' of innovation? *Technovation, 14*(2), 93-110.

Fukugawa, N. (2006). Science parks in Japan and their value-added contributions to new technology-based firms. *International Journal of Industrial Organization, 24*(2), 381-400.

Furukawa, R. & Goto, A. (2006). Core scientists and innovation in Japanese electronics companies. *Scientometrics, 68*(2), 227-240.

Gibbons, M. et al., (1994). *The new production of knowledge. The dynamics of science and research in contemporary societies*. London/Thousand Oaks/New Delhi: SAGE Publications.

Gulbrandsen, M. & Smeby, J.-C. (2005). Industry funding and university professors' research performance. *Research Policy, 34*(6), 932-950.

Gurney, T. et al., (2012). Author disambiguation using multi-aspect similarity indicators. *Scientometrics, 91*(2), 435-449.

Hansson, F. et al., (2005). Second generation science parks: from structural holes jockeys to social capital catalysts of the knowledge society. *Technovation, 25*(9), 1039-1049.

Hekkert, M.P. et al., (2007). Functions of innovation systems: A new approach for analysing technological change. *Technological Forecasting and Social Change, 74*(4), 413-432.

Horlings, E. & Gurney, T., (2012). Search strategies along the academic lifecycle. *Scientometrics*, 1-24.

Jones-Evans, D. (1995). A typology of technology-based entrepreneurs: a model based on previous occupational background. *International Journal of Entrepreneurial Behaviour & Research, 1*(1), 26-47.

Klofsten, M. (1999). Supporting the pre-commercialization stages of technology-based firms: the effects of small-scale venture capital. *Venture Capital: An International Journal of Entrepreneurial Finance, 1*(1), 83-93.

Koh, F.C.C. et al., (2005). An analytical framework for science parks and technology districts with an application to Singapore. *Journal of Business Venturing, 20*(2), 217-239.

Lanciano-Morandat, C. et al., (2009). Le capital social des entrepreneurs comme indice de l'émergence de clusters ? *Revue d'économie industrielle*(4), 177-205.

Langrish, J. et al., (1972). *Wealth from knowledge: a study of innovation in industry*: Halstead Press Division, Wiley.

Lepori, B. et al., (2007a). Comparing the evolution of national research policies: what patterns of change? *Science and Public Policy, 34*(6), 372-388.

Lepori, B. et al., (2007b). Indicators for comparative analysis of public project funding: concepts, implementation and evaluation. *Research Evaluation, 16*(4), 243-255.

Löfsten, H. & Lindelöf, P., (2005). R&D networks and product innovation patterns--academic and non-academic new technology-based firms on Science Parks. *Technovation, 25*(9), 1025-1037.

Magerman, T. et al., (2010). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics, 82*.

Markman, G.D. et al., (2005a). Innovation speed: Transferring university technology to market. *Research Policy, 34*(7), 1058-1075.

Markman, G.D. et al., (2005b). Entrepreneurship and university-based technology transfer. *Journal of Business Venturing, 20*(2), 241-263.

Meyer, M. (2002). Tracing knowledge flows in innovation systems. *Scientometrics, 54*(2), 193-212.

Moed, H.F. et al. (Eds.),. (2004). *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems.* Dordrecht: Kluwer Academic Publishers.

Mowery, D.C. et al., (2001). The growth of patenting and licensing by US universities: an assessment of the effects of the Bayh-Dole act of 1980. *Research Policy, 30*(1), 99-119.

Mustar, P. et al., (2006). Conceptualising the heterogeneity of research-based spin-offs: A multi-dimensional taxonomy. *Research Policy, 35*(2), 289-308.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization science, 5*(1), 14-37.

Phillimore, J. (1999). Beyond the linear view of innovation in science park evaluation An analysis of Western Australian Technology Park. *Technovation, 19*(11), 673-680.

Quintas, P. et al., (1992). Academic-industry links and innovation: questioning the science park model. *Technovation, 12*(3), 161-175.

Siegel, D.S. et al., (2003). Assessing the impact of university science parks on research productivity: exploratory firm-level evidence from the United Kingdom. *International Journal of Industrial Organization, 21*(9), 1357-1369.

Song, Y. et al., (2007). *Efficient topic-based unsupervised name disambiguation*. JCDL'07, Vancouver. ACM, New York.

Steffensen, M. et al., (2000). Spin-offs from research centers at a research university. *Journal of Business Venturing, 15*(1), 93-111.

Stokes, D.E. (1997). *Pasteur's Quadrant: Basic Science and Technological Innovation*. Washington, D.C.: Brookings Institution Press.

Tijssen, R.J.W. (2002). Science dependence of technologies: evidence from inventions and their inventors. *Research policy, 31*(4), 509-526.

van den Besselaar, P. & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: a method and an exploratory case study. *Scientometrics, 68*(3).

Volberda, H.W. et al., (2010). Perspective - absorbing the concept of absorptive capacity: How to realize its potential in the organization field. *Organization science, 21*(4), 931-951.

Westhead, P. & Batstone, S. (1998). Independent Technology-based Firms: The Perceived Benefits of a Science Park Location. *Urban Studies, 35*(12), 2197-2219.

Westhead, P. & Storey, D.J. (1995). Links between higher education institutions and high technology firms. *Omega, 23*(4), 345-360.

Zahra, S.A. & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. *Academy of Management Review, 27*(2) 185-203.

Zucker, L.G. & Darby, M.R. (1996). Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry. *Proceedings of the National Academy of Sciences of the United States of America, 93*(23), 12709.

2 Author Disambiguation Using Multi-Aspect Similarity Indicators[1]

Abstract

The key to accurate bibliometric analyses is the ability to correctly link individuals to their corpus of work, with an optimal balance between precision and recall. We have developed an algorithm that performed this disambiguation task with a very high recall and precision. The method addresses the issues of discarded records due to null data fields and their resultant effect on recall, precision and F-measure results. We have implemented a dynamic approach to similarity calculations based on all available data fields. We have also included differences in author contribution and age difference between publications, both of which have meaningful effects on overall similarity measurements, resulting in significantly higher recall and precision of returned records. The results are presented from a test data set of heterogeneous catalysis publications. Results demonstrate significantly high average F-measure scores and substantial improvements on previous and stand-alone techniques.

2.1 Introduction

The use of scientometrics has become increasingly prevalent in many forms of scientific analysis and policy-making. Key to good bibliometric analysis is the ability to correctly link individuals to their respective corpus of work, with an optimal balance between precision and recall when querying the larger data set in which their corpus resides. This is especially important where bibliometrics is used for evaluation purposes. The most common problem encountered is that of multiple persons having the same last name and initial. Other problems include misspelled names, name abbreviations and name variants. Within a small data set, these errors can be corrected using manual checks. However, when data sets are large, time and labour constraints severely hamper disambiguation efforts. The increasing scale and scope of scientometric studies (Cassiman et al., 2007; Moed et al., 2004; Phelan, 1999; Trajtenberg et al., 2006) and the rapid rise of Asian science systems - where there is substantially lower variance in names - reinforce the need for an automated approach to author disambiguation.

There is a need for algorithms designed to extract patterns of similarity from different variables, patterns that can set one author apart from his or her namesake, and link to other data sources. Our primary focus in this paper is the problem of correctly identifying multiple persons sharing the same last name and first or all initial or initials. We have developed a novel algorithm that increases the precision and recall of author-specific records, whilst decreasing the number of records discarded due to missing data. The algorithm allows for factors such as author contribution, time difference between publications and dynamic combinations of indicators used.

1 A shortened version of this chapter is published as Gurney, T. et al., (2012a). Author disambiguation using multi-aspect similarity indicators. *Scientometrics, 91*(2), 435-449.

The paper is structured as follows. In the next section, we review the literature on the disambiguation methods that are currently being employed, with an emphasis on methods that mix both computer science, and sociological and linguistic approaches. We pay particular attention to the prevalence of data discarding and the effect it has on results, leading to the data and method section. The data and method section presents an overview of the strategy of the method, expectations of current methods (including our own), the data used, and data preparation. We explain in detail which metadata objects are used and how these objects are employed in similarity calculations and logistic regression. We also explain how our method utilises two new metadata objects - time between publications and author contributions - to achieve stronger disambiguation results. In the results section we apply our method to a test data set to show its accuracy in terms of precision and recall. In the final section, we discuss the results and look forward to the next generation of methods for disambiguation.

2.2 Previous work

The current literature on disambiguation is split between computer science and sociological and linguistic approaches. Few papers have brought the two approaches together, which seems the more fruitful approach. They are discussed briefly in this section.

Zhu et al. (2009) constructed string and term-similarity graphs between authors, based on publication titles. Graph-based similarity and random walk models were applied with reasonable success to data from DBLP. A similar study by Tan et al. (2006) uses search engine result co-occurrence for author disambiguation. Yang et al. (2008) discovered disambiguation problems in citations and developed a method to determine correct author citation names using topic similarity and web correlation with the latter providing stronger disambiguation power. Kang et al. (2009) also use co-author web-based correlations and co-author-of-co-author (co-author expansion) techniques in their study. In disambiguating researcher names in patents, Raffo and Luillery (2009) investigate the different search heuristics and devise sequential filters to increase the effectiveness of their disambiguation algorithm. Song et al. (2007) have developed a two-stage approach to assist with the problem of disambiguating persons with the same names on web pages and scientific publications; first using two topic-based models -Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA)- linking authors to words and topics, and then using a clustering method -hierarchical agglomerative clustering- to disambiguate names. The testing was conducted using web data and CiteSeer on a data set of 750,000 papers.

The problem of ambiguity is also addressed in studies dealing with heterogeneous data sets. For example, by linking patents to publications and authors to inventors (Cassiman et al., 2007; Meyer, 2001; Raffo & Lhuillery, 2009; Trajtenberg et al., 2006).

The most relevant recent publications come from Tang and Walsh (2010) and Onodera et al., (2011). Using the concept of cognitive maps and approximate structural equivalence, Tang and Walsh developed an algorithm based on the *knowledge homogeneity* characteristics of authors. They analysed the effectiveness of their technique on two common names (one of English origin, the other of Chinese origin). Their technique was remarkably successful, but potentially biased in that records that did not exhibit any similarities in the cited references were treated as isolates and were consequently excluded from the final effectiveness results. The study by Onodera et al.

is the most similar to ours in that they use similarity probabilistic techniques. They differ not only in their objective of disambiguation (as their aim is to retrieve specific authors' documents, not to discriminate between different authors within a data set) but also by their use of the available metadata fields. The fields they used include co-authorships, affiliation addresses, citation relationships, title words, interval in years between publications, author country, citation and co-citation relationships.

2.3 Data selection and discarding of records

The discarding of data seems to be a prevalent feature of the existing literature. In many cases, where only one information object is used in the analysis, it is logical that if that field is empty or null, the records cannot be used in the analysis. This results in discarded data and a loss of recall, and the lost data is not always addressed. Some examples of data discarding are discussed here.

Malin (2005) uses the peripheral social network of actors to disambiguate specific entities. Presumably most movies have at least one actor so this method may be sound for the particular data source. However, to translate this to the scientific sphere where co-authorships are used to disambiguate researchers, publications often have a single author and by discarding all publications with only one author is a drastic move. A recent study by Kang et al. (2009) use co-authorships in publications to disambiguate but make no mention of the issue of single-author papers in their methodology. Huang et al. (2006) rely on author metadata as input for similarity calculations. However, the metadata examples given are emails and URLS, and addresses and affiliations. The problems with the selection of this sort of information include authors email addresses generally only including the corresponding author's email address, and physical addresses typically not being author-specific (where, for example, there are 5 authors but only 2 addresses are given).

2.4 Data and Method

The objective is to create a network of publication/author nodes in which edge strengths are the probabilistic value of the two nodes being the same person, as calculated by logistic regression. A community detection algorithm is employed over the network to discriminate the pairings of nodes in terms of unique authorship.

2.4.1 Realistic expectations of disambiguation techniques

Techniques for author disambiguation are based on the assumption that the source data - whilst not providing a unique identifier for every author - as a minimum will spell the author's last and first names correctly. This assumption has been proven to be naive in almost all data repositories, as there are multiple avenues for error to creep in. However, if one endeavours to correctly assign a whole corpus of works to one person, using all the misspelled variants of a person's name would dramatically increase the effort required to minimally increase the recall of that particular author's work. Furthermore, databases such as the Web of Science mobilise authors to correct metadata, such as correcting misspelled authors' names. As authors have a vested interest in correctly spelled names, one may expect this type of mistake to be increasingly resolved.[2] As a

2 However, corrections are not always possible, partly due to the structure of a database. This holds for entries older than 1995 (email 3 March 2011, from Thomson Reuters)

result of this, we have chosen to discard any variants in the spelling of the last name, and will rely on one spelling of the name.

2.4.2 Data

In the testing and implementation phases of this project we have used heterogeneous catalysis data collected by a project team within the PRIME ERA Dynamics project. The data set is a collection of 4979 articles, letters, notes and reviews featuring 5616 authors. The records were retrieved from Thomson Reuters' Web of Science (WoS) and parsed using SAINT (Somers et al., 2009). Through manual cleaning and checking, each publication was assigned to the correct author. Each record is considered unique, and is based on a combination of the article and author IDs assigned during the parsing process. There are 3872 different last names and of these there are 2014 last names which have more than one publication. There are 4403 author last name and first initial variants, with 208 instances in which more than one author has the same last name and first initial and 366 authors who share their last name only with one or more other authors. We have focused our efforts on the instances in which there are more than one author with the same last name.

2.4.3 Data preparation

Each author/publication combination was assigned a unique identifier (U ID). This is to ensure that each and every author instance is regarded as unique at the beginning of the process. The contingent of metadata present in each publication, and associated U ID, were marked. We have selected the following base metadata from the available metadata of WoS and provide an explanation for the choice and for the treatment of potential problems of the metadata:

1. Publication title words: title word choice by authors is generally considered to be related to content. Assuming the author is relatively consistent in his field, content (and thus title word choice) will remain relatively stable (Han et al., 2003) and the relative level of co-occurrence of title words between publications gives a strong indication of whether Author *A1* is the same as Author *A2*. However, this may not be constant across fields or in fields with stylised titles. The changing lexicon and meaning of words may also play a part. Also, title words may have been chosen specifically to address a particular audience - the so-called "audience effect" (Leydesdorff, L., 1989; Whittaker et al., 1989).

2. Publication abstract words: As with publication titles, word choice is related to content along with perceived application benefits and a general overview of the methodology and results. The additional data of application benefits and methodology gives a more detailed picture of the cognitive background of the work, which in turn gives more depth of information to the similarity comparison algorithm. With both title and abstract words, we removed stop words and stemmed words using SAINT (Somers et al., 2009).

3. Citations: Working within the same field, a researcher may base much of their his or her on specific previous studies in the field, adding to the unique 'characterisation' of their work. Citation behaviour is also punctuated by levels of self-citations, group citations and opportunistic citation (Aksnes, 2003; Nicolaisen, 2007; Pasterkamp et al., 2007) which only add to the characterisation of the citation list. It is this behaviour that allows citations to be regarded as

an indicator of similarity. However, citations not only suffer from ambiguity themselves, but citation behaviour may be different between fields and therefore differently contribute to identification through similarity. We have chosen to use citations 'as is' and have not manually checked ambiguous citations.

4. Keywords: A publication generally contains both author-generated keywords and journal-indexer-generated keywords which can be used to create a measure of similarity between two publications (Matsuo & Ishizuka, 2004). Author-generated keywords may be more accurate reflections of the content rather than the indexers' keywords due to the "indexer effect" (Healey et al., 1986). Keywords (or more accurately, 'key-phrases') are normalised by removing spaces between words and by grouping highly similar key-phrases based on Damerau-Levenshtein edit distances (Damerau, 1964; Levenshtein, 1966).

5. Author listings: Researchers tend to co-author within their own field, generating co-authorship lists that do not diverge enormously from their home field. Co-authorship occurrences are not necessarily field-dependent and when researchers do co-author, they tend to do so repeatedly within the same topic areas (Wagner-Döbler, 2001). The higher the shared co-author count across different publications, the higher the likelihood that authors with the same name are indeed the same individuals. Co-author names are used in a 'last name, first initial' format as not all records maintain a listing of all the authors' full names.

6. Author addresses: Addresses are commonly used in disambiguation studies as they may definitively link an author to an address and if two authors of the same name share an address the likelihood that they are the same person is high. However, the use of addresses is complicated by authors maintaining more than one address (guest lectureships etc), by inconsistent spelling of addresses, incomplete addresses, no address given, or when multiple authors and multiple addresses exist on publication data (Tang & Walsh, 2010)[3]. The addresses are normalised for object order (for example - house number followed by street name versus street name followed by house number) by using Damerau-Levenshtein distances. In the case of multiple addresses and multiple authors with no defined indication of author-address links, a probabilistic approach is used where each author on a publication has an equal probability of linking to any of the addresses presented.

7. Journal name: Research fields may be delineated by the set of core journals in which most publications are published. Assuming a level of consistency in researchers' chosen fields, the primary choices of which journal to publish in remain relatively constant (van den Besselaar & Leydesdorff, 1996). However, changing journals in a field and any inter/multi//transdisciplinary research output may not be targeted to a constant list of journals, resulting in a lower degree of similarity when comparing author publications (Loet Leydesdorff et al., 1994).

---

3 Various data repositories, principally WoS, are working to improve the issue of multiple assigned author addresses, and newer publications in the database have direct indications of which author/authors link/links to which address/addresses.

To complement the metadata fields given, our similarity calculations use additional data. These are:

1. Difference in years between publications: The age difference between publications will have an effect on the degree of similarity between publications as there may be a change in the individual's research focus over time, and with that, a change in popular co-authors, choice of title words and/or keywords and so on. Allowing for this time difference may also change the role played by the base metadata in discerning the probability of two publications being by the same author.

2. Average author contribution: With indicators such as publication title, publication abstract, citations, and choice of journal - the selection of words, citations and journal is performed in various but typically unequal measures by the contributors in the publication (Bates et al., 2004; Yank & Rennie, 1999). Therefore, it is necessary when using the indicators to take this inequality in contribution into account. For example, if a researcher is listed as 3rd or 4th author, the probability that he or she has contributed heavily to word choice in the title or citations is lower than if he had been 1st or 2nd author on the publication. Author contributions are calculated using the sum of the fractional author counts of the author positions of the two records using Moed's formula (2000). The contributions of the second and last authors are equal to 2/3 of the contribution of the first author. Any other authors contribute 1/3 of the first author. This is normalised so that the sum of all the fractions is equal to one[4]. For example, in a publication of 6 people, where *a* is the contribution of the first author:

   $$a + 2/3a + 1/3a + 1/3a + 1/3a + 2/3a = 1;\ and\ a = 3/10$$ (Moed, 2000).

   The author of a single-authored publication has maximum control over input, and from the formula of Moed - $a=1$ and thus the maximum value for contribution is 1. The average author contribution measures the deviation from maximum input, i.e. how 'far' away an author is from the maximum. For each author pair being compared, the average distance from maximum of each author is the average author contribution (AAC), and this is on a scale of 0-1, where 1 signifies maximum input of the two authors being compared, that is to say that both authors are the only author in their respective publications.

2.4.4 Null combination code (NC)

When each record is compared, the minimum shared available metadata of each pair is referred to as the Null Combination (NC) code. This "null data field code" (NC) is a string of ascending order numbers where each digit signifies the presence of a valid field. For example, if only the title, labelled as "1", abstract -"2" and author assigned keywords -"4" are present the NC code will be 124.

4 In the case of alphabetical listings of authors, each author is assigned a value of 1/n (where n is the total number of authors).

2.4.5 Year difference (YD) categories

The YD is categorised as follows: 1) ≥ 2 years difference; 2) >2 and ≤ 5 years difference; 3) >5 and ≤ 10 years difference; 4) >10 years difference

2.4.6 Similarity calculations

The similarity calculations are based on the Jaccard index, which is calculated on the following metadata fields:[5]:

1) title words; 2) abstract words; 3) last names and first initials of co-authors; 4) cited references in whole-string form; 5) normalised author keywords; 6) normalised indexer keywords; 7) normalised research addresses; 8) journal names.

2.4.7 Logistic Regression

Logistic regression requires the presence of two predetermined groups. We start by identifying some of the authors' correct publications and some publications with the same author name that definitively belong to another group. With this, we created an input data set in which the predetermined groups are defined as Group 2 (where the author/publication records being compared are definitively the same individuals) and Group 0 (where the author/publication records being compared are definitively NOT the same individuals) as shown in Table 1.[6]

Table 1 Sample input data table for regression analysis

	Pairs		Independent variables			
Group	A	B	1	2	3	NC Code
0	1	5	a	b	c	123
0	1	6	NULL	b	c	23
0	…	…	…	…	…	…
2	2	3	NULL	b	c	23
2	1	3	NULL	b	NULL	2
2	..	…	…	…	…	…

Rathenau Instituut

The independent variable (the metadata fields) cells contain the raw similarity values of those independent variables. If the independent variable is not present to compare between U IDs, it is marked as being NULL. The NC code reflects which of the independent variables are present for each U ID pairing. The data was split into calibration and testing sets in an approximately 25:75 ratio to test the validity of the model. A regression was run with the NC codes as filters.

5 The metadata fields are compared across records that share the same last name only.

6 To make the algorithm useful for completely unchecked sets and thus avoid excessive manual checking of records we are currently working on a sampling method which will be presented in a follow-up publication.

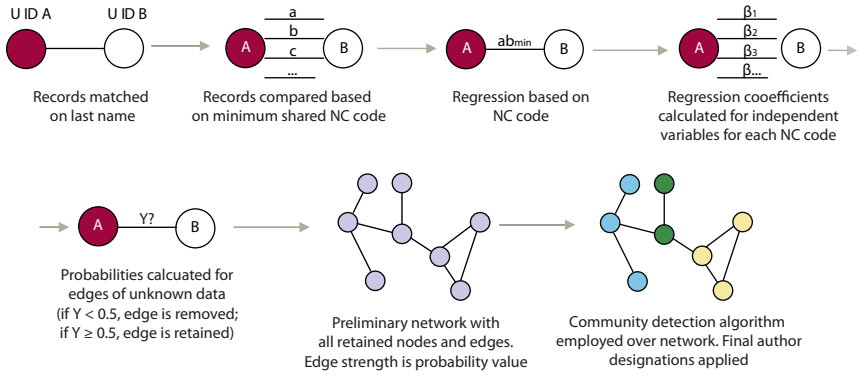
The full regression formula is as shown in Equation 1. For each NC combination (where "Sim" refers to the degree of similarity between two author/publication pairs in a specific type of metadata):

Equation 1

$$\ln(\mathbf{Y}/1\text{-}\mathbf{Y}) = \beta 0 + \beta_1(\text{SimCoauth}) + \beta_2(\text{SimAbstract}) + \beta_3(\text{SimTitle}) + \beta_4(\text{SimCitedRef}) + \beta(\text{SimAuthorKeywords}) + \beta_6(\text{SimIndexerKeywords}) + \beta_7(\text{SimRes.Address}) + \beta_8(\text{SimJournal}) + \beta_8(\text{AAC}) + \beta_8(\text{YDCategory})$$

The β coefficients found in the regression are used to estimate the pairing probabilities of the unknown data set. The default decision rule threshold of .5 is used to determine calculated group membership. The flowchart in Figure 1 summarises the order of operations in which the calculations are performed.

Figure 1 Summary of order of operations of data processing, regression calculations and final author disambiguation



2.4.8 Final Author Assignment

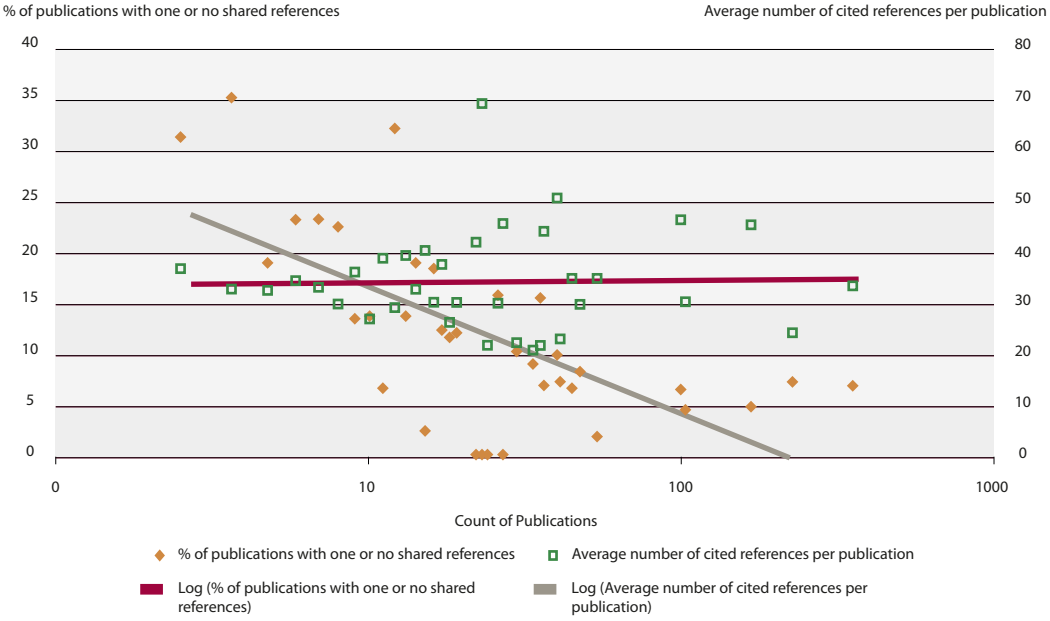
Final author designation is performed by the community detection algorithm of Blondel et al. (2008). This algorithm takes into account the weighted edges of a network and assigns each node to a specific community based on the surrounding nodes and their edge weights. Logistic regression predicts the probability of two publications being from the same author on a row by row basis, but the community detection algorithm works on the entire interconnected network of nodes or publications and identifies the communities of papers belonging to unique authors.

2.5 Results

To demonstrate the effectiveness of our algorithm we have chosen to present results based on matching last names only, and on matching last name and first initial. To demonstrate the

importance of average author contributions and the time difference between publications, we present the β values of the logistic regression for these variables from each NC code. The authors we have chosen as examples are all of the authors from our data set who have more than one publication and whose last name is shared by one or more other authors. This list consists of 366 different authors with publication counts ranging from 2 to 420.

Figure 2 Average percentage of publications with one or no shared cited references compared to average number of cited references per publication



Rathenau Instituut

As a precursor to our results, we take a critical look at the potential problems of other methods, by using our own data of 366 different authors to demonstrate some of the fallibilities of alternative approaches. We have chosen to use the cited references as many previous studies use only cited references in their disambiguation efforts. We tested the number of cited reference matches between records and compared this to the mean number of cited references per publication. We have plotted the percentage of records which have one or no shared cited references and this is presented in Figure 2.

In Figure 2, as per the trend line, the mean number of cited references per publication does not deviate depending on how many publications an author has written. Contrary to this, the number of publications that share one or no cited references does vary with the number of total publications an author has written: the more publications an author has written, the fewer the publications that share one or no cited references. When disambiguating authors with relatively few publications, there is a much higher chance that the recall of their publications will be affected

because there are fewer shared cited references. The scarcity of shared cited references may have a negative impact on precision. This may be due to the exploratory nature of 'young' scientists' work. This is an important statistic to take into account when considering cited references as the primary source of metadata for disambiguation. It is difficult to build up a picture of the characterising aspects of an author if there are few or no similar characteristics between his or her publications.

For our primary results, we have used the harmonic mean version of the F-measure, with equal emphasis placed on precision and recall (Do et al., 2009). The F-measure (equation *2c*) is composed of the precision and recall values as shown in equations 2a and 2b:

Equation 2a

Precision (P) = (TruePositive)/(TruePositive+FalsePositive)

Equation 2b

Recall (R) = (TruePositive)/(FalseNegative+TruePositive)

Equation 2c

F-measure = 2*(PR/P+R)

We have calculated the average precision, recall and F-measure values on authors with varying counts of publications, by using their last name, and last name and first initial.

2.5.1 Contributions of AAC and YD

Table 2 shows the β coefficients of AAC and YD to the logistic regression calculations.

Table 2 Contributions to group membership in logistic regression calculations by AAC and YD

NC Code	YD β	AAC β
357	0.14	2.484
1357	0.2	2.551
2357	0.073	1.784
3457	0.184	2.411
12357	0.125	1.911
13457	0.22	2.481
13579	0.154	1.128
23457	0.1	1.766

NC Code	YD β	AAC β
23579	0.128	1.277
34578	0.102	1.828
34579	0.128	0.97
123457	0.137	1.873
123579	0.174	1.292
134578	0.14	1.887
134579	0.166	1.055
234578	0.107	1.644

NC Code	YD β	AAC β
234579	0.146	1.212
345789	0.134	0.856
1234578	0.142	1.722
1234579	0.182	1.236
1345789	0.17	0.881
2345789	0.153	1.096
12345789	0.186	1.065

Rathenau Instituut

Note: The NC code signifies the available metadata objects. The presence of each number signifies the presence of a specific metadata object. The numerical codes for each object are: 1: Co-authorship, 2: Abstract, 3: Title, 4: cited References, 5: Journal, 7: Research Address, 8: Author Keywords, 9: Indexer Keywords, (6: Journal Category is not shown)

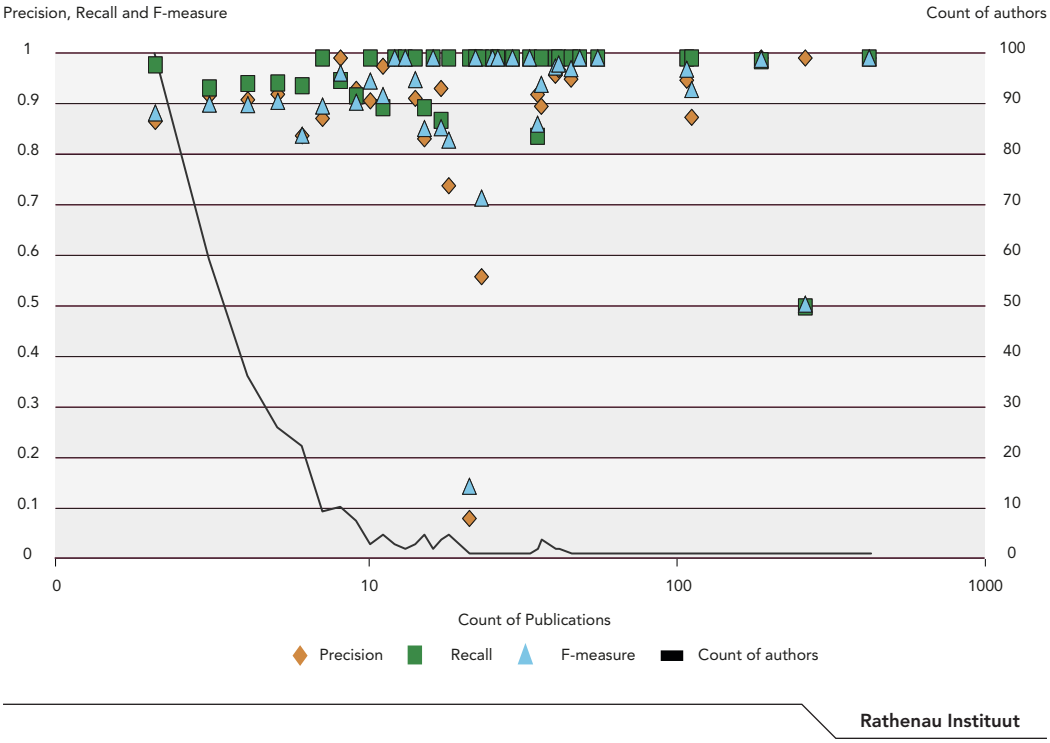
For every NC code the AAC β is always higher than the YD β. The maximum possible value of the AAC is 1, signifying that the closer the two authors are to having maximum input on the publication, the higher the chances that the edge between the two publications in the network will be regarded as being a correct edge, i.e. the two publications are by the same person.

The further away the two authors are from maximum input, the lower the chances that the edge will be placed between the two publications. Of the variables available, the indexer keywords are not affected by the authors in any way. The research addresses are also not affected by the authors themselves as they are indicators of location rather than content. The variability of the AAC β when one takes into account exactly what input authors have on a publication is something to be investigated in the future. For YD, the β coefficients do not vary much over the different NC codes, which was unexpected as we hypothesised that the effect of time difference on similarity between publications would affect results more significantly. The results indicate that the effect of time difference between publications decreases when the abstract is included in the analysis. Abstracts seem to have a larger similarity over time through a recurring use of some words.

2.5.2 Results based on last name only

Figure 3 shows the average precision, recall and F-measure of authors of varying publication counts. The distribution of the number of authors with specific counts of publications is as expected, i.e. there are many authors who have published little, and few authors that have published prolifically. The average recall values per publication total are all above 0.85.

Figure 3 Average recall, precision and F-measure values including author publication count distribution, using last name only



The precision values are mostly higher than 0.9 with a few exceptions. Almost all the F-measure values are above 0.8 with most above 0.9, with one exception at 0.5.

The exceptions to these scores are primarily due to two different authors with the same last name but different first initial being incorrectly designated as a single author, in which one of the authors has an exceptionally high count of publications (~200) and the other a relatively low count of publications (~20). A similar situation which affected the F-measure occurred when one author was deemed to belong to two communities, i.e. the algorithm has classed one author as two separate authors. Where this occurred, we used the average of the "two" authors to give a single result.

Overall, the results based on last name only are very high, which constitutes a very good result, considering the number of authors, and the count of authors with the same last name.

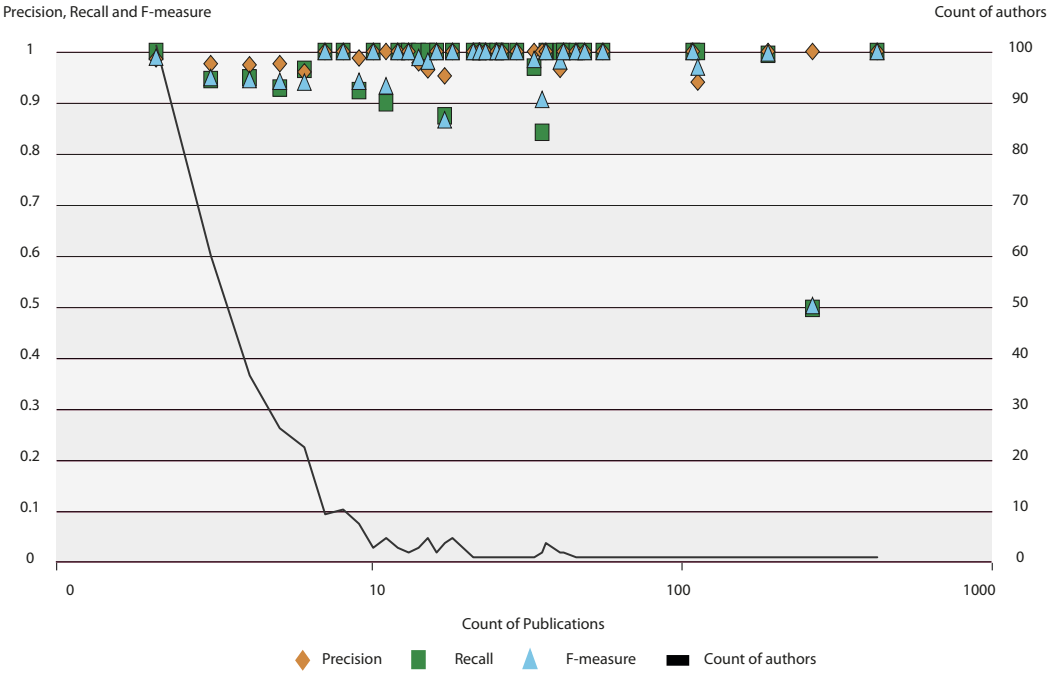
2.5.3 Results based on last name and first initial

Figure 4 shows the average precision, recall and F-measures of the same authors which were presented in Figure 3, but now on a last name and first initial level. The distribution of number of authors with specific publication counts remains the same. Compared to Figure 3, the results by last name and first initial are better in almost all aspects, as expected. There is no longer the

problem of two authors with the same last name but different first initials being incorrectly classed as the same author. The low result at 0.5 in both figures is from one author who has been incorrectly designated as two separate authors. On overview of Figure 4, the results for each publication count category of authors, has increased as compared to Figure 3. There are many more perfect scores (precision=1, recall=1, F-measure=1) over many more publication count categories.

To summarise the results as seen in Figures 3 and 4, the algorithm worked remarkably well across the range of authors with different publication counts. This is important as it shows that the algorithm is extremely suitable for discerning those authors who have few publications and who may not publish repeatedly in the same field. An example of this may be young scientists' PhD-related publications, as compared to their post-doctoral or further publications.

Figure 4 Average recall, precision and F-measure values including author publication count distribution by last name and first initial



For authors who have many publications, the algorithm works well to assign these publications to the correct author, who may have changed research topics multiple times over the years.

2.6 Discussion and conclusions

Author disambiguation will continue to pose a problem for some time, even with database providers working to solve the problem. The move towards placing the onus of identification on

authors may be a step forward. But the records of authors who are no longer active in publishing may remain ambiguous for the foreseeable future. It is for this reason that algorithms such as ours will remain important for researchers who make use of bibliometric data.

Our decision to compare records by the last name and first initial, and by last name only, was a result of our need to test the algorithm's discerning power and robustness. By creating a very large number of possible matches (our final master table of potential match records had over 1.5 million rows) we intended to stress our computing power, and the ability of the algorithm to handle such a large number of records.

Our method differs from previous methods in three ways. Firstly, we do not preselect records based on specific metadata. Rather, we utilise every available metadata object. Retaining and utilising all possible metadata has proven to be helpful, because records that did not display any similarity on cited references, for instance, may still have shown high topic similarities as seen in the abstract or title words. A key fact of the data we examined was that there were a substantial proportion of records that were missing metadata. The variability of what metadata was available to compare spurred us to think of a dynamic approach which would only need to use the minimum shared metadata. This meant that a record could be compared to others by completely different metadata for each comparison. The use of logistic regression was required for this as we wanted to know the contributions of each data object to discerning group membership, and we realised that for each combination of available data objects, there would be different levels of contribution by each object.

Secondly, the additional metadata that we have chosen to include was also important: time difference between publications and average author contribution. The goal of our disambiguation method (and that of many other similar methods) is to create a continuous chain of publications - a coherent sub-network within the larger network. Allowing for the age difference between publications increased the chances of linking young publications to older ones, rather than just linking similarly-aged publications to each other. Changing topics of, and influences on, a researcher over time create a longitudinally stretched network of publications which, when thresholds are applied, are susceptible to being broken. Linking the older and younger publications increases the likelihood of that sub-network remaining intact. The average author contribution metadata was very important as it gave the algorithm scope for flexible similarity parameters. Tang and Walsh (2010) mention the fact that other authors in a publication have an influence on what metadata is included in the final version of the publication, thus affecting the "knowledge homogeneity" of the author under inspection. We have successfully shown that recognising and, more importantly, using this difference in author contribution actually increases the coherence of the sub-network of publications by a specific author. Together, these two additions to the range of employed metadata increase the deductive power of the algorithm.

The retention of all possible metadata has also proven to be helpful as records that did not display any similarity on one variable, such as the cited references, may still have shown high topic similarities in other variables, such as the abstract or title words. More importantly, an author's contribution to each publication ultimately affects what title words, abstract words, and cited references etc are used. This is a very important factor when considering similarity-based disambiguation methods such as ours.

Previous studies commonly use thresholds to increase accuracy rates, which are useful in a proof-of-concept, but in real situations there is no way to know which threshold is the best to use. Our method does not use any thresholds, apart from the default .5 threshold for logistic regression which, when applied to real-world operations, is far easier to manage and replicate for further studies.

To move our algorithm from proof-of-concept to a working process, we need to address the issue of pre-checking records. There is a substantial amount of manual work involved in all methods (including ours). At present, excluding the previous authenticity checks performed by the originators of the data set, the method - from parsing publications to final author designation - takes approximately 8 hours, of which the most time is spent importing the logistic regression results from SPSS into Access. The use of a plug-in for R (an alternative statistical analysis program) is being investigated which would vastly reduce the time spent.

A drawback of this method surfaces when individuals publish in multiple, unrelated fields. Unless there are bridging publications that exhibit similarities to more than one distinct publishing field, the networking aspect will show separate clusters, thus affecting precision and recall. With the benefit of further research, we will investigate the minimum number of publications necessary to consistently and accurately disambiguate authors.

To summarise, our method retains all data and discards no information, accounts for activity of authors in different fields or specialties (year difference) and in different capacities (AAC), uses no arbitrary thresholds, is scalable, and provides highly accurate disambiguation results.

This algorithm and technique could be further applied to most forms of entity resolution, such as that of inventors and applicants in the patenting field. We hope to develop it in such a form soon.

Author ambiguity is a serious enough issue to warrant more attention. We hope that through our method we will be able to improve upon past efforts and to eventually present a user-friendly, open-source tool for scientists, policy-makers and evaluators, so that decisions based on error prone results become less common. We aim to integrate this disambiguation tool into SAINT (available from reference website). This would allow records from various data repositories to be parsed and accurately sorted by author or inventor to the order of hundreds of thousands of records.

2.7 References

Aksnes, D.W. (2003). A macro study of self-citation. *Scientometrics, 56*(2), 235-246.

Bates, T. et al. (2004). Authorship criteria and disclosure of contributions: comparison of 3 general medical journals with different author contribution forms. *Jama, 292*(1), 86.

Blondel, V.D. et al. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*, P10008.

Cassiman, B. et al. (2007). Measuring industry-science links through inventor-author relations: A profiling methodology. *Scientometrics, 70*(2), 379-391.

Damerau, F.J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM, 7*(3), 171-176.

Do, H.H. et al. (2009). Comparison of schema matching evaluations. *Web, Web-Services, and Database Systems*, 221-237.

Han, H. et al. (2003). *A model-based k-means algorithm for name disambiguation*. ISWC2003, Florida 2003.

Healey, P. et al. (1986). An experiment in science mapping for research planning. *Research Policy, 15*(5), 233-251.

Huang, J. et al. (2006). Efficient name disambiguation for large-scale databases. *Lecture Notes in Computer Science, 4213*, 536.

Kang, I. S. et al. (2009). On co-authorship for author disambiguation. *Information Processing & Management, 45*(1), 84-97.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(8), 707-710.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy, 18*(4), 209-223.

Leydesdorff, L. et al. (1994). Tracking areas of strategic importance using scientometric journal mappings. *Research Policy, 23*(2), 217-229.

Malin, B. (2005). *Unsupervised name disambiguation via social network similarity*. SIAM International Conference on Data Mining. Newport Beach 2005.

Matsuo, Y. & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools, 13*(1), 157-170.

Meyer, M.S. (2001). Patent citation analysis in a novel field of technology: An exploration of nano-science and nano-technology. *Scientometrics, 51*(1), 163-183.

Moed, H.F. (2000). Bibliometric indicators reflect publication and management strategies. *Scientometrics, 47*(2), 323-346.

Moed, H.F. et al.(Eds). (2004) *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems.* Dordrecht: Kluwer Academic Publisher

Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology, 41*(1), 609-641.

Onodera, N. et al. (2011). A method for eliminating articles by homonymous authors from the large number of articles retrieved by author search. *Journal of the American Society for Information Science and Technology, 62*(4), 23.

Pasterkamp, G. et al. (2007). Citation frequency: A biased measure of research impact significantly influenced by the geographical origin of research articles. *Scientometrics, 70*(1), 153-165.

Phelan, T.J. (1999). A compendium of issues for citation analysis. *Scientometrics, 45*(1), 117-136.

Raffo, J. & Lhuillery, S. (2009b). How to play the "names game": Patent retrieval comparing different heuristics. *Research Policy, 38*(10), 1617-1627.

Somers, A. et al. (2009). Science Assessment Integrated Network Toolkit (SAINT): A scientometric toolbox for analyzing knowledge dynamics. The Hague: Rathenau Instituut.

Song, Y. et al., (2007). *Efficient topic-based unsupervised name disambiguation*. JCDL'07, Vancouver. ACM, New York.

Tan, Y.F. et al. (2006). *Search engine driven author disambiguation*. Paper presented at the 6th ACM/IEEE-CS joint conference on Digital libraries, Chapel Hill, NC, USA.

Tang, L. & Walsh, J.P. (2010). Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics, 84*(3), 763-784.

Trajtenberg, M. et al. (2006). The Names Game: Harnessing Inventors' Patent Data for Economic Research. *NBER working paper*.

van den Besselaar, P. & Leydesdorff, L. (1996). Mapping change in scientific specialities; a scientometric case study of the development of artificial intelligence. *Journal of the American Society of Information Science, 47*(5).

Wagner-Döbler, R. (2001). Continuity and discontinuity of collaboration behaviour since 1800 - from a bibliometric point of view. *Scientometrics, 52*(3), 503-517.

Whittaker, J. et al. (1989). Creativity and conformity in science: titles, keywords and co-word analysis. *Social Studies of Science, 19*(3), 473-496.

Yang, K.H. et al. (2008). Author Name Disambiguation for Citations Using Topic and Web Correlation. *Research and Advanced Technology for Digital Libraries*, 185-196.

Yank, V. & Rennie, D. (1999). Disclosure of researcher contributions: a study of original research articles in The Lancet. *Annals of internal medicine, 130*(8), 661.

Zhu, J. et al. (2009). A Term-Based Driven Clustering Approach for Name Disambiguation. *Advances in Data and Web Management*, 320-331.

3 Analysing Knowledge Capture Mechanisms: Methods and a Stylised Bioventure Case Study[1]

Abstract

Knowledge transfer between science and technology has been studied at micro and macro-levels of analysis. This has contributed to the understanding of the mechanisms and drivers, but actual transfer mechanisms and processes, be they through codified or tacit sources, have very rarely been mapped and measured to completeness and, to a large extent, remain a black box. We develop a novel method for mapping science-technology flows and introduce 'concept clusters' as an instrument to do so. Using patent and publication data, we quantitatively and visually demonstrate the flows of knowledge between academia and industry. We examine the roles of exogenous and endogenous knowledge sources, and of co-inventors and co-authors in the application of university-generated knowledge. When applied to a stylised case study, we show that the method is able to trace the linkages between base knowledge and skill sets and their application to a technology, which in some instances span over twenty-five years.

3.1 Introduction

Knowledge transfer between universities and firms has become increasingly institutionalised (Geuna & Muscio, 2009) as universities look for novel, more insightful, ways to enhance their economic and societal value through new technology spin-offs or start-ups (Audretsch et al., 2005; Tijssen, R.J.W., 2006). Much of the previous literature has focused on the facilitating actions and conditions for knowledge transfer such as scientific publications, conferences, informal interactions, collaborative and contract research, IP licensing, personnel exchanges and hiring - each with varying significance for industry (Ponomariov & Boardman, 2012).

A major challenge to evaluating these knowledge transfer routes and mechanisms is uncovering meaningful linkages between technological outputs and scientific inputs. Knowledge transfer occurs most often at both the codified and tacit level, and the transfer processes and motivations within academic research versus those in industry settings are complex and evolving. However, what is not discussed in detail in the existing literature is the demarcation and measurement of the knowledge that is transferred (Bozeman, 2000). This is of utmost importance because the facilitation of transfer has been investigated but the question of *whether* knowledge has been transferred can only be answered by (a) being able to demarcate the object of transfer, and (b) measuring its point of inception, evolutionary path and eventual application.

1 This chapter has been published as Gurney, T. et al. (2012). *Knowledge Capture Mechanisms in Bioventure Corporations.* Proceedings of Science and Technology Indicators (STI), Montreal and has been submitted to The Journal of Informetrics.

Specific quanta of knowledge evolve along developmental paths, shaped not only by the scientific and technological developments of the laboratory in which it was conceived, but also by the further learning and skill sets of the scientists and inventors involved. By exploring the routes of codified knowledge transfer from inception to exploitation, we can begin to understand the processes and mechanisms of knowledge transfer. These include interactive knowledge production, the role a scientist's skill set plays, the effect of a scientist's peers - be they in the university or in the lab - and the transformative nature of science itself.

There have been substantive efforts to examine the facilitation processes and end-utilisation in an isolated sense, analysing each step in the overall development process of a technology. However, a novel methodological approach is necessary to address the question of whether the whole transfer process has occurred. As mentioned previously, this requires a method to both demarcate and track specific quanta of knowledge. Doing this grants us a clear view on the effectiveness of the facilitating conditions.

In this paper we use and adapt available tools and models, and integrate those with newly developed tools to provide a complete picture of knowledge transfer, from start to finish. This paper starts with a discussion of the role of the scientist/entrepreneur, and that of his surroundings, in developing the necessary skill sets and knowledge for eventual transfer and application in industry. We then apply these insights to our methodology, which is described in detail. A crucial step in the methodology is the introduction of the idea of 'concept clusters', which refers to a small, cognitively cohesive agglomeration of scientific peer-reviewed publications. As an illustration, we briefly apply the methodology to a case study. In the conclusions, we summarise the potential benefits, open methodological issues, and routes for further research.

3.2 Conceptual framework

The codification of knowledge takes two primary forms: patents and scientific publications. The use of patents as indicators was pioneered by Schmookler (1966), followed by many applications (such as Schmoch (1993) and Fleming (2001)). However, many aspects of their indicator-orientated uses do have drawbacks (Pavitt, 1988). For example, not all innovations are patented (Arundel, 2001; Arundel & Kabla, 1998), with many innovations kept under a veil of secrecy (Brouwer & Kleinknecht, 1999), leading to underestimation of innovative potential or capacity. Analyses using patent indicators are typically based on metadata found in patents. Title words, abstract words and keywords (Courtial et al., 1993; Engelsman & van Raan, 1994), patent classifications (Leydesdorff, 2008; Tijssen, R.J.W. & Van Raan, 1994), and patent/non-patent citations (Karki, 1997; Meyer, M.S., 2001) have all been used extensively. Many patent databases exist from which we extract the metadata used in analyses, each with their own idiosyncratic advantages and disadvantages. These include disclosure requirements of prior art ('duty of disclosure'): the USPTO requires an exhaustive list but the EPO requires a minimal listing. Differences also stem from the databases themselves, in terms of their formatting, whilst others relate to the practices of applying for patents through different national or supranational patenting offices. Despite the stated shortcomings, patents can be used for mapping knowledge transfer in a large part of the knowledge-intensive economy because patent documents are highly detailed descriptions of the processes, applications and necessary information required for a technology. Citations within a patent document, either to other patent documents or scientific

literature, add to this wealth of data. Patent documents encompass a wide range of technological fields and the major patenting offices (such as the USPTO or EPO) cover patent data from all countries (Tijssen, 2001).

Publications serve as the primary indicators for the defining characteristics and development of science. They are the most visible outcome of scientific endeavours, with an extensive range of indicators and methodologies developed. The analysis of publications shares a number of analytical approaches with patent analyses, such as word mapping (Callon et al., 1991) and citation analysis (Garfield & Welljams-Dorof, 1992; White & McCain, 1998). Using co-occurrences of combinations of words and cited references in publications is also becoming a common technique (Braam et al., 1991; van den Besselaar & Heimeriks, 2006).

The act of publishing itself is subject to a complex system of social and scientific norms, practices and reward systems (Merton, 1957). Publishing behaviours and patterns of scientists are governed in large part by these norms and practices, as well as by serendipity. The development of a university scientist's profile and portfolio are the result of search strategies (Horlings & Gurney, 2012) employed by the scientist. University-based scientists publish primarily to extend their professional and intellectual prowess, and regular publishing is considered a requirement. Industry-based scientists are governed by similar constraints, and the firm benefits from publishing too - by becoming intimately involved with the basic science behind the technologies (Rosenberg, 1990), and their publications serve as a signal of their capabilities to the outside world (Hicks, 1995).

The conditions required for facilitating the development and transfer of knowledge depends heavily on the recipient knowledge platform. Knowledge assets (Nonaka, 1994), sector roles (Baba et al., 2009) and science-push and demand-pull concepts (Langrish et al., 1972), are all factors in a knowledge base's receptivity. In this manner - external knowledge sources, taking into account demand and current capabilities, are readily absorbed and entrained into stock knowledge bases and practices. This receptivity is known as 'absorptive capacity' (Cohen & Levinthal, 1990) and can best be described as "[t]he ability of a firm to recognize the value of new, external information, assimilate it, and apply it to commercial ends is critical to its innovative capabilities," (p.128). The individuals involved are at the heart of this, with the absorptive capacity of a firm tied to its constituent individuals' absorptive capacity, i.e. the right personnel are in place to take advantage of incoming information. As Cohen and Levinthal (1990) state, "Beyond diverse knowledge structures, the sort of knowledge that individuals should possess to enhance organizational absorptive capacity is also important. Critical knowledge does not simply include substantive, technical knowledge; it also includes awareness of where useful complementary expertise resides within and outside the organization" (p.133).

The use of patent and publication data, in the context of absorptive capacity, allows us to map knowledge inputs and outputs, and consequently illuminate the mechanisms at work. The aim of this study is to provide a map of the cognitive route between the scientific origins and the technological output, with specific focus on the knowledge capture mechanisms operating. To this end, we have developed a method that shows:

1. *How the scientific background of the patent corpus links to the scientific output of the inventor.*
   A patented product is the result of accretion over time of the research results, practices, skill sets and processes of the inventors involved. In the patent documents one can identify references to the underlying science (cited publications) and technology (cited patents) that were instrumental towards the development of the new (patented or non-patented) technology. Linking the patent corpus and publication output allows us to determine the background or necessary scientific requirements for the technologies.

2. *How the collaborative research environment of the researcher/inventor contributes to the development of the underlying science and to the technology developed.*
   Academic and industrial collaboration is common in high-technology fields. Much of science is the result of collaborative efforts between researchers, where resources can be pooled and task allocation increases efficiency. As such, any contributions from a researcher's network will be visible in any publication authorship list or patent inventor list.

3. *What other knowledge is needed by the researcher/inventor for the development of a technology, and how this is appropriated.*
   Scientists must incorporate new results and skills from previous research done by others, to improve upon and modify their own intellectual prowess and breadth of skills. Their individual absorptive capacity of the individual is measured by their entrance into, and adoption and integration of, new fields cited by the technologies they work in.

By mapping these three aspects of the knowledge stream, we can clarify several of the mechanisms through which knowledge capture is supported: (1) the researcher/inventor's own research; (2) the researcher/inventor's collaboration network; (3) the researcher/inventor's knowledge uptake process.

In order to map aspect 1, we have developed an approach based on the overlap in content of the non-patent literature references (NPLRs) found in the patent applications, and the publication corpus of the inventor. An individual publishes in multiple streams of research, with the streams being composed of publications highly similar to each other, which can be determined algorithmically. The similarity between the researcher/inventor's publications and the NPLRs can be calculated so that the NPLRs are co-located together with the research streams of the individual.

Comparing the underlying total knowledge and skill set required to develop the technology with the knowledge and skills of the individual researcher/inventor shows the contribution of the latter to the technologies. The contributions of an individual's co-inventors and co-authors can be similarly constructed allowing us to map aspect 2. Finally, in order to map aspect 3, a more refined approach is required. An individual's research streams may be broad in topic and time, and general statements can be made regarding the relevance and importance of an individual's knowledge and skills to a company's technologies. In order to examine the specific scientific fields that the technology draws upon (as defined by the NPLR and the fields from which they originate), we have developed a method focusing on the specific scientific concepts and methods necessary for the technologies described in the patent documents. By identifying the specific concepts utilised in the technologies and the point in time that the researcher/inventor develops or integrates them into his or her knowledge base, we are able to view from where, and from

what original form he or she derived new knowledge assets. This method utilises *concept clusters*, which will be defined and operationalised in detail in the next section. Concept clusters are used to map aspect 3 and to examine the detailed concepts and methods in aspects 1 and 2.

3.2.1 Concept clusters

A broad description of an individual's knowledge and skill sets may be derived through examination of the titles used, references cited, keywords used (and more) in their publication corpus. Adding the NPLRs of the patent applications to the individual's publication corpus allows us to discern which aspects of an individual's corpus are similar to the NPLR. To discern general research themes within the combined corpus, we utilise the Louvain clustering method (Blondel et al., 2008) which optimises the modularity of a network, i.e. the actual distribution of edges between nodes versus a random distribution, to identify macro-clusters in the network. The metadata occurrences in each cluster are then examined to identify the general themes. To identify *specific* topics, each macro-cluster is isolated and the same clustering algorithm is applied to produce micro-clusters. These micro-clusters constitute the immediate environment of the NPLRs. Depending on the variety of subjects in the publication corpus, macro-clusters can range in size from 10 to 100+ publications whereas each micro-cluster is typically no larger than 10 publications.

We refer to these micro-clusters or immediate environments as 'concept clusters'. The publications cited by the patent applications (NPLRs) make up the nucleus of the concept cluster and each concept cluster contains at least one NPLR. Surrounding this nucleus are the publications most similar in terms of title word and cited reference combinations (van den Besselaar & Heimeriks, 2006), and the borders of each concept cluster are algorithmically delineated into communities (Blondel et al., 2008). A concept cluster contains, in varying proportions, publications authored by the researcher/inventor (which the patent applications may or may not cite), and publications written by others that are cited by the patent application. The specific composition of a concept cluster describes the knowledge utilised in the patent application, in terms of the knowledge base and skills internal or external to the researcher/inventor.

Table 1 Concept cluster composition

		Publication authored by	
		Inventor	Other
Cited by patent	Yes	A	B
	No	C	-

Table 1 illustrates the possible publication origin types - A, B and C - in a concept cluster. For each concept cluster, a mix of publication type can result. If the concept cluster contains:

1. Type A publications - this indicates direct contributions by the inventor to the required concepts and skill sets. The research and concepts contained within the publication are either necessary for, or directly related to, the development of the technology.
2. Type B publications - we assume that some knowledge is outside the expertise of the inventor.
3. Type C publications - whilst the inventor is not cited directly, his or her publications are highly similar to publications that are being cited. The inventor's skill sets and background knowledge are similar to the NPLR.

As defined previously, absorptive capacity is the ability to recognise, assimilate and integrate new knowledge, and apply it in a novel manner. The absorptive capacity of an individual who is an inventor of the technologies can be determined by analysing the similarity and presence (or lack thereof) of their contributions to the concept clusters. The greater the number of occurrences of their own publications that are similar to the NPLR, the higher the enabling potential for absorptive capacity.

The timing of an inventor's publishing entry into a concept cluster is important. If publications by an inventor appear (Type A or C) first, followed by NPLRs (Type B), we conclude that the inventor has already previously developed the skill sets and knowledge required for the technologies. If NPLRs appear first and are then followed by the inventor's publications, we conclude the inventor previously did not have the skills or knowledge necessary but has had to address this. How soon an inventor publishes after recognising the NPLR indicates the perceived importance of that knowledge to those technologies, and to the inventor's own knowledge stock. This increases the similarity between the NPLR and the IA's publications, and consequently the absorptive capacity.

Using this approach, we can map aspect 3 from the previous section, and add it to the first two. We determine whether an inventor is a leader in the production of the knowledge required for their technologies, or a follower. If a follower, does the inventor incorporate the necessary knowledge and skill sets into their portfolio early, demonstrating a high level of absorptive capacity? If necessary knowledge and skills lie outside the portfolio of the inventor, do his collaborators provide any of the knowledge or skills?

3.3 Previous work

Previous studies typically utilise text-mining approaches or citation matching to provide a linkage between patents and publications. Text-mining approaches generally involve methodologies that identify topical clusters in patents and publications using words (title, abstract, or full text) and link the two corpora together through the similarities between the topical clusters. Mogoutov et al. (2008) use a combinatorial approach to map innovation in the biomedical field of microarrays. Relevant concepts are extracted from multiple data sets, namely those of publications, patents, and research project data. A matching algorithm links the data sets through their shared concepts. They specifically try to avoid using pre-determined topic areas or research areas, to allow some qualitative room for interpretation after the matching has been completed. They successfully demonstrate a link between, and within, scientific fields through shared concepts.

Magerman et al. (2010) provide a very thorough review of the state of the art of the text-mining approach. In addition, their study tests the effectiveness of distance measures when linking patents and publications via text mining. With only 30 patents and 437 publications, Magerman et al. use a smaller data set than what is typically encountered. These are notable figures, because commonly-used similarity distance measures rely on large data sets to provide higher-quality matching outcomes. The authors acknowledge this and conclude that the overall number of records would likely increase the chance of linking patents and publications.

Text mining can rely on an abundance of methods, which are highly variable and customisable. However, some limitations of text mining also become apparent. The different vocabularies employed between patents and publications pose a threat to accurate matching. The size of the sample may result in misleading or inaccurate matching options. A further limitation is one of a changing vocabulary over time within a field of science. In publishing, the audience and indexer effects (Leydesdorff, 1989; Whittaker, 1989) may lead to fewer and fewer matches between publications and patents further apart in time. Text mining is typically a resource-intensive approach, and requires extreme care due to the complex nature of linguistic behaviours and anomalies.

Citation matching is easier as it involves extracting the bibliographic non-patent literature references (B-NPLRs) from the patent documents and finding the corresponding twin in which-ever publication database one uses. Unfortunately, the requirements for including citations (patent and non-patent) in patent applications vary drastically between patenting offices, making the duty of disclosure a prime example relevant to this study. The move to include in-text non-patent literature references (IT-NPLRs) is a recent development as the availability of extraction tools for full-text documents has increased. A study by Tamada (2006) addresses the issue of IT-NPLRs, focusing specifically on Japanese patent documents. They argue that as there is no requirement by the Japanese Patent Office to include front-page references, patent output indicators that utilise only B-NPLR may miss relevant scientific references. To counter this, they use references found in the text of the documents to successfully identify under-reported scientific fields cited by patent applications. They conclude that the inclusion of both in-text and bibliographic citations enriched their data sets and provided balance between objective and strategic referencing of literature in patent applications.

Meyer (2002) examined the use of citations in patent and publication-centred studies. He formed a typology of the most frequently used approaches, such as patent citation analyses, industrial scientific activities, and university and academic patenting. His critiques of the techniques essentially point to the misuse of analytical tools and methods from one field to another. He notes that techniques that use these approaches do not take certain fundamental basic characteristics of patenting and publishing into account. For instance: firstly, different fields show a different propensity to publish; secondly, citations can be negative or positive; and thirdly, publishing is not the only output of the laboratory. In terms of patenting, similar problem characteristics should be taken into account, such as: the patenting propensity varies across industries, not all inventions are patented, and a significant proportion of patents are strategic, designed to block innovation by a competitor. Meyer may have examined these aspects over ten years ago, but the principles remain valid today when discussing methodologies using citation behaviours of publications and patents. Regarding NPLRs in patents, the relative abundance of

references to scientific literature versus non-scientific literature is an indicator of the quality (Branstetter, 2005) and proximity to science (Callaert et al., 2006) of the patent application. What is generally understood and accepted is that placing citations to scientific literature in patent documents indicates a cognitive link to, or awareness of, the related scientific concepts (Tijssen, R.J.W., 2001).

3.4 Data and Method

The methodology we developed consists of various steps. The first step is to select the inventor/researchers that play a crucial role in the relevant knowledge transfer case study. The second step is data collection of the papers and patents of these individuals (4.1). Then we do publication clustering (4.2) and patent application clustering (4.3). The next step is to link the patent applications and publication clusters, using a specific visualisation tool (4.4). After having described the method, we demonstrate it in section five: proof of concept.

3.4.1 Data

There are many patent databases around, all with their own idiosyncrasies, some of which stem from the databases themselves whilst others relate to the practices of various national or supranational patenting offices. In our study we use the PatSTAT database prepared and developed by the EPO, as it aggregates various other databases, and is considered one of the most extensive patent databases. For our publication data, we use the Thomson Reuters' Web of Science (WoS) as our primary source of publication data, supplemented by CV data from the scientists involved. The sources and types of data come from:

1. Patent data - we extracted all patent applications with the selected inventors each listed as an applicant from the EPO PatSTAT database along with all other inventors' data; this also included all patent applications with the firm under study listed as an inventor or assignee, and the selected inventors as assignees.

2. Publication data[2] - we extracted from WoS all publications with the inventors' firm listed as an institution; and all publications with the selected inventor listed as author.

These base data were parsed using SAINT[3](2009) and managed in a relational database. Further data were collected from the patents. More specifically, these were (and where they were found):

1. Bibliographic NPLRs (B-NPLRs) - these are citations included primarily by the examiner and added as front-page references.

2. In-text non-patent literature references (IT-NPLRs) - citations to publications visible in the body of the patent. These IT-NPLRs were automatically extracted from the full-text versions of the patent documents by custom software.

---

2 English language only

3 SAINT (Science-system Assessment Integrated Network Toolkit - a Rathenau Instituut open-source software suite designed to parse, clean and organise bibliometric data to be used later in relational database software such as MS Access and MySQL.

All patent applications were then grouped according to their first filing, with the priority patent application representing the collective. Single-priority based families are collections of patent applications that claim a specific application as the first or priority application. The priority patent is included in the collection (Martinez, 2010). This is done to account for variations in NPLR reporting and inclusion across different patenting offices. A second reasoning is that any derivative applications are close extensions of the priority patent, thus one could expect the NPLRs from the collective to extend to the other applications in the group. Further references in this paper to these patent collectives use the term 'priority patent' to mean the *priority patent application representing the collective*.

The NPLRs were then normalised for search of their twin in WoS. If there were no NPLRs linked to any given priority patent, the NPLRs of derived citing patent applications (i.e. applications citing the original application as priority) were included. Both NPLR sets were parsed and, as far as possible, their WoS publication equivalents found. A manual check was performed to see if the retrieved documents matched the original NPLR. If any discrepancies in metadata did not allow for a proper match, the affected records were not utilised in any further analysis. The verified documents were then parsed and processed together with the inventor's publications to create a master publication database and the origins of each document were recorded.

3.4.2 Publication similarity and concept clusters

Publications are clustered by their shared combinations of title words and cited (van den Besselaar & Heimeriks, 2006). The degree of similarity is calculated using the Jaccard similarity coefficient. Clusters of publications were automatically assigned by a community detection algorithm (Blondel et al., 2008) grouping publications based on their degree centrality and relative weights of edges between nodes.[4] These clusters are referenced further as 'research streams'. Each research stream is then isolated and the community detection algorithm of Blondel et al. is run on the individual streams resulting in the concept clusters.

3.4.3 Patent clustering

Patent applications are grouped by INPADOC family ID and the NPLRs of the INPADOC families within concept clusters are noted. The Jaccard similarity coefficient is calculated between INPADOC families using the shared concept clusters in which their NPLRs occur, and the community detection algorithm of Blondel et al. is used to designate INPADOC clusters.

3.4.4 Visualising patents and publications

We have developed a method (Horlings & Gurney, 2012) that allows the specific research trails that an individual has developed to be visualised in a uniquely clear manner. We have built upon this method by adding patent applications whose individual researchers are listed as an inventor to their corpus of publications. The thematic and knowledge base aspects of the patents and publications are linked, not through direct citations by patents to the publications, but through

4 For a more detailed explanation of clustering algorithms in general, see Palla, G. et al., (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*(7043), 814-818. For a comparative analysis of Blondel et al.'s algorithm versus others' see Lancichinetti, A. & Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical Review E, 80*(5), 056117.

shared thematic or topical research areas of the cited NPLRs and the inventors' corpora of publications. Even if the patent document does not cite the individual's publication corpus directly, other cited literature may (or may not) cluster within the inventor's areas of expertise. This approach results in a tangible, visible, shared knowledge base between the patent and the publication.

We arranged the patent applications and research streams along two axes: time on the x axis, research streams and patents on the y axis. Longitude is defined as [(year of publication) - (year of first documented publication in the data set)/(range in years) x 360] - 180. Please note that the small clusters of papers visible within each research stream represent the annual production within that stream, and not the concept clusters - the latter contain papers published over multiple years. Latitude is defined as [(stream number)/(total number of streams) x 180] - 90. The nodes were positioned with the GeoLayout in Gephi (2009), using an equirectangular projection.

3.5 Proof of concept

In this section we apply the methodology to a stylised case study - stylised, as we are primarily interested in demonstrating that the methodology is able to map the knowledge streams as well as the mechanisms underlying these streams. This is based on a real case study that we aim to analyse in depth elsewhere (Gurney et al., 2013). The essential mechanics of the methodology are discussed here, with additional detail provided in the in-depth study. Images here are stylised, with data utilised to illustrate the necessary components.

3.5.1 Case study selection

Our case study involves a prominent biotechnology researcher who is strongly involved in cancer therapeutics at the firm he founded in 2001 and the university at which he is a professor. (The individual, firm and university shall further be referred to as IA, FA and UA respectively.) IA maintains direct links between his research at UA and research conducted at FA. This enables us to draw upon his extensive publishing history as well as his numerous patenting activities at both university and firm.

3.5.2 Data summary

Table 2 shows a summary of the various data collected. Patents cover the period 2000-2008[5], and the publications cover all publications in the categories defined in the previous section up to 2011. The large number of patent applications (242) may not be typical of most companies in this field. The breadth of the patent applications, as exemplified by the number of INPADOC families (90) is also large. IA is a prolific author with, in 2011, 931 publications to his name. This is an exceptionally high number and we assume many of these publications are purely the result of him being head of a large institute in which his name appears as author as a matter of seniority.

5 Patent applications up to 2008 were chosen as there is considered to be a delay in the completeness of patent data in PatSTAT. 2008 was chosen as the last year as we could be more certain that all possible patent data was included.

Table 2 Summary of collected patent and publication data of IA

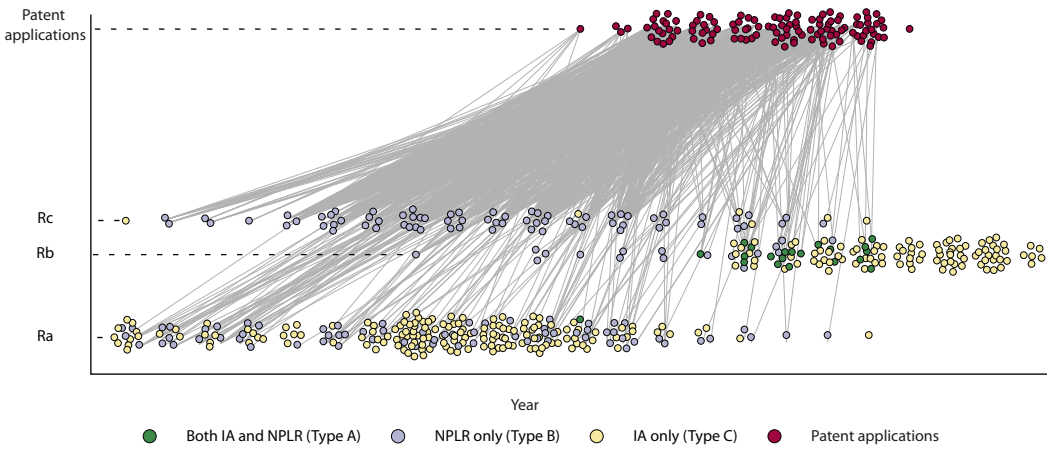
Data	Feature	Count
Patent Applications	Patent applications with FA listed as assignee and IA as inventor (2000-2008)	242 (115 priority patent applications)
	INPADOC families	90
Publications	IA publications retrieved from WoS	931 (786 pre-2009)

Rathenau Instituut

1. *Mapping the links of the scientific background of the patent corpus to the scientific output of the inventor.*

Figure 1 is a stylised image showing portions of the total corpus of IA publications and the NPLRs of the patent applications differentiated into research streams. Noted in Figure 1 are the publications authored by IA and/or publications cited by the patent applications over time. The visualisation was constructed as detailed in section 4.4. Research stream ***Rb*** contains a considerable number of NPLRs authored by IA as seen by the black nodes. Streams **Ra** and **Rc** contain NPLRs not authored by IA, publications authored by IA but not cited and very few NPLRs authored by IA co-located in the same stream. Each stream may contain a mixture of publication types (A, B or C), and the proportional presence of IA's publications (cited or not) in the stream indicates the proximity of IA's research to the research cited by the patent applications.

Figure 1 Stylised image of patent applications and publications - authored by IA and/or cited by the patent applications over time



Rathenau Instituut

Note: For descriptions of type A,B,C nodes - see table 1

In Table 3, the NPLR distribution in the patent applications is noted. Most of the NPLRs come from within the text of the patent documents. 65 NPLRs are found in both the text and bibliography of the applications. IA's publications make up 10% of the NPLRs, with a proportionally larger number being cited in the bibliography.

Table 3. Summary of NPLR

Data	Feature	Count
NPLRs	Unique NPLRs found (and matched to WoS) in all patent applications	525[a] unique (2037 total)
	In B-NPLR only	147[b]
	In IT-NPLR only	313[c]
	In both B-NPLR & IT-NPLR	65[d]
NPLR citations	Count of IA's publications cited in NPLR (Note: % = x/a)	55 (10%)
	In B-NPLR only (%=x/b)	18 (12%)
	In IT-NPLR only (%=x/c)	19 (6%)
	In both B-NPLR & IT-NPLR (%=x/d)	18 (27%)

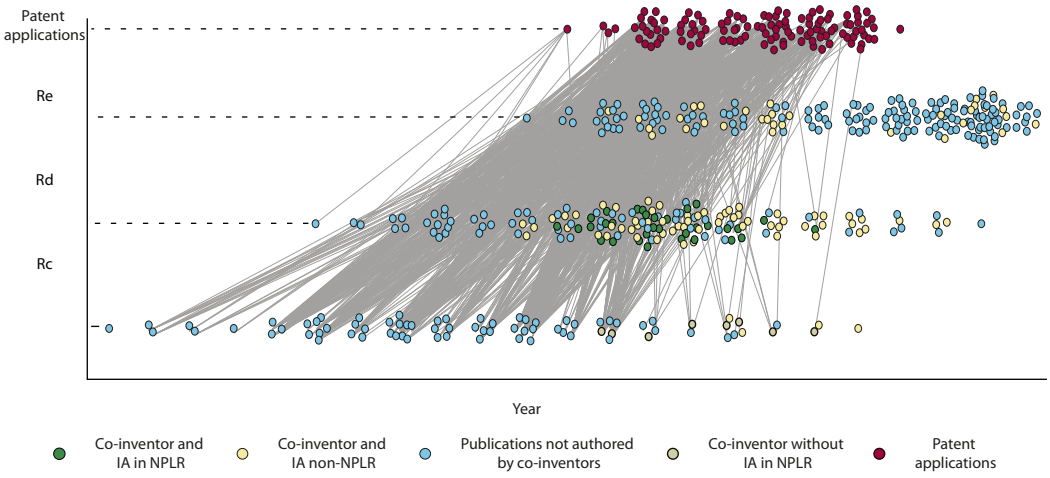
Rathenau Instituut

We can conclude from Figure 1 and Table 3 that aspects of research conducted by IA are relevant and necessary to the technologies represented by the patent applications. Some aspects are not within IA's expertise, such as those in research stream **Rc**. IA's research is cited in many instances in both the text of the document and the bibliography. Research that is cited but not authored by IA is often very similar to IA's publications, as seen in stream **Ra**.

2. *Mapping contributions of the inventor's research collaborations in the patent and publication data.*

Figure 2 is a stylised image of portions of the total corpus of IA publications and NPLRs. The presence of co-inventors as authors of publications is noted. In research stream **Rc**, there are a number of NPLRs authored by IA's co-inventors but do not feature IA as an author. In stream **R*d*** there is a high proportion of NPLRs and non-NPLRs authored by both IA and his co-inventors. Stream **Re** contains many publications that are authored by both IA and his co-inventors but not cited by any of the patent applications. All three streams contain many papers by IA which were not co-authored by the co-inventors, and these papers are partly NPLRs, and partly non-NPLRs. The streams also contain NPLRs not authored by IA or his co-inventors.

Figure 2 Co-inventors of IA and their presence in NPLR and publications



Rathenau Instituut

Table 4 summarises the presence of co-inventors' publications as NPLRs in the whole corpus. Only 9 publications written by IA's co-inventors (without IA as author) are cited by the patent applications in the NPLRs (8 are shown in stream **Rc** in Figure 2, the last is located in another stream) and most are IT-NPLR (in-text). The number of co-inventors' publications cited by patent applications (excluding publications co-authored by IA) is far lower than the number of IA's publications cited by the applications. IA's co-inventors appear as inventors without IA on 30 patent applications.

Table 4 Summary of co-inventors' publication and patent application contributions

	Category	Feature	Count
Co-inventors	Publications	Cited in NPLR (excluding publications co-authored by IA)	9
		In B-NPLR only (%=x/b)	2 (1.5%)
		In IT-NPLR only (%=x/c)	5 (1.5%)
		In both B-NPLR & IT-NPLR (%=x/d)	2 (9%)
		Publications not cited by patent applications butco-authored with IA	251
	Patent applications	FA patent applications without IA as inventor	30 (12%)

Rathenau Instituut

Note: for b, c and d values see Table 3

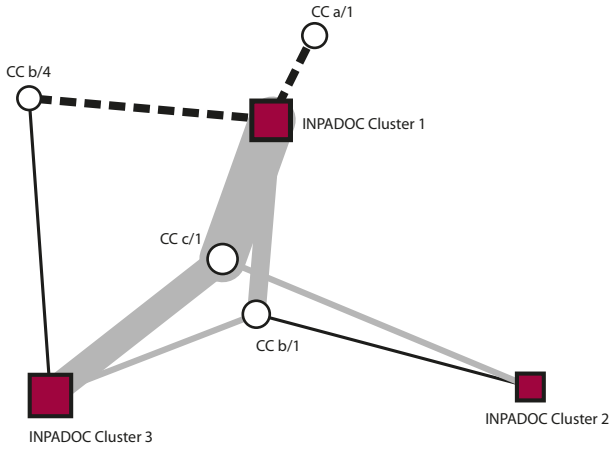
3. *Mapping the inventor's level of adaptive knowledge use (absorptive capacity) necessary for the development of a technology.*

We map the concept clusters and their utilisation by patent applications over time, in order to demonstrate the absorptive capacity of IA in relation to the development of the technologies. To construct the concept clusters, each research stream in the total corpus is isolated. The community detection algorithm of Blondel (2008) is run over each isolated stream. Each resulting community or concept cluster contains a mixture of publications (Types A, B and/or C from Table 1), with at least one NPLR forming a nucleus. We grouped each patent application into INPADOC clusters, based on the similarities of the aggregated main group IPC codes of the patent applications' parent INPADOC families. The resulting INPADOC clusters represent the different technologies in which IA is involved, and their development over time.

Figures 3-5 show stylised representations of the appearance in time of the concept clusters containing NPLRs cited by the INPADOC clusters. The concept clusters represent not only the immediate knowledge environment of the NPLRs, but also the degree of similarity between the inherent skill sets of IA and the skill sets referenced by the patent applications in the NPLRs. This allows us to map the knowledge and skill sets that are necessary for the development of the technology (and therefore cited by the patents). IA's own publications may not be cited in many instances but are highly similar to the NPLRs and are co-located in the same concept cluster. This mapping strategy also shows the stage of the technologies' development at which the skill sets and knowledge from outside IA's expertise are cited.

Figure 3 includes concept clusters that contain only NPLRs not authored by IA (Type B from Table 1). The skill sets and knowledge base contained in these publications are considered to be outside the expertise and skill sets of IA as they do not contain any similar IA publications (in terms of title words or cited references).

Figure 3 Concept clusters cited by INPADOC clusters containing only NPLRs not authored by IA (Type B publications)

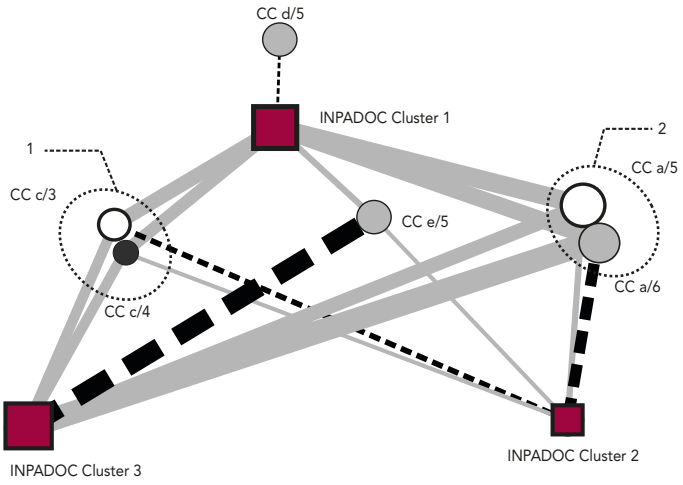


Rathenau Instituut

Note: Concept clusters are labelled CC. Size of nodes indicates count of publications and count of INPADOC families. Thickness of edges indicates number of citing INPADOC families. Edge colour indicates at what phase in the age of the INPADOC cluster the concept is cited, grey=early, dashed=middle, black=late.

As is visible in Figure 3, the NPLRs located in the concept clusters are cited at different stages in the development of the technologies of INPADOC clusters 1-3. For example, CC b/4 (a concept cluster found in research stream **Rb**) is cited in the middle development period by INPADOC cluster 1, but at a later period by INPADOC cluster 3 whereas CC c/1 (a concept cluster found in research stream **Rc**) is cited early in the development of all the INPADOC clusters. CC a/1 from **Ra**, is cited in the middle development phase by only INPADOC cluster 1.

Figure 4 Concept clusters cited by INPADOC clusters containing NPLRs not authored by IA (Type B) & IA publications not cited by patent applications (Type C)



Rathenau Instituut

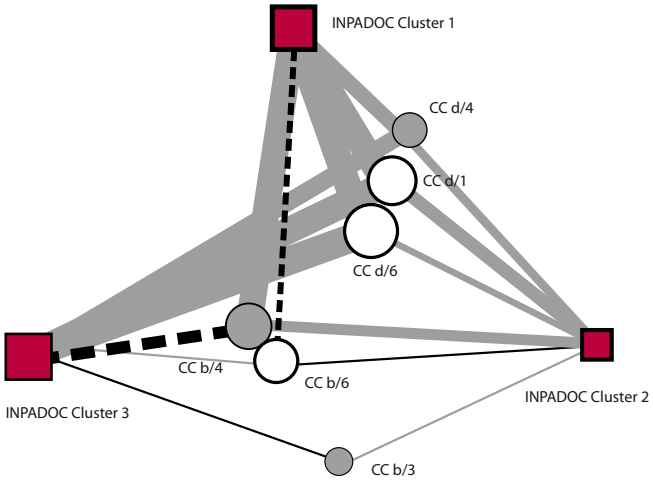
Note: Concept clusters are labelled CC. Node size indicates count of publications and count of INPADOC families. Node shading indicates time of appearance of IA's publication into concept cluster, white=early, grey=middle, black=late. Edge thickness indicates number of citing INPADOC families in the INPADOC cluster. Edge colour indicates period of citation, grey=early, dashed=middle, black=late.

Figure 4 shows concept clusters containing a mixture of NPLRs not authored by IA (Type B) and publications authored by IA but not cited by the patent applications (Type C). In the highlighted area 1 are two concept clusters (CC c/3 and c/4), both from research stream **Rc**. These are cited in the early phases of development of INPADOC clusters 1, 2 and to a lesser extent 3, and CC c/3 is cited in the middle phase of development by INPADOC cluster 2. IA's publications appear early in CC c/3 but late in CC c/4. This implies that the knowledge and skill sets of CC c/3 are necessary from an early period and IA is publishing in this concept cluster also at an early stage, whereas IA only starts publishing in CC c/4 at a late stage in the concept cluster's lifespan. In highlighted area 2, concept clusters CC a/5 and a/6 from research stream **Ra** are cited in the early to middle development phase of the technologies. In CC a/5, IA's publications are found early in the lifespan of the concept cluster, but in CC a/6, his publications only appear in the middle stages of the concept cluster's lifespan.

Relating this to IA's absorptive capacity: IA has already published similar publications taking into account the macro-clusters (considering CC c/3 & c/4 and CC a/5 & a/6 are from the same respective research streams) and highly similar publications considering each concept cluster

separately. For the concept clusters in which IA's publications appear only in the middle or late phases, we can conclude that IA has recognised the importance of the research being cited by the patent applications and, demonstrating a degree of absorptive capacity, begins to populate the concept clusters with his own publications (not cited by patents).

Figure 5 Concept clusters cited by INPADOC clusters containing NPLR authored by IA (Type A), NPLR not authored by IA (Type B) and IA-authored publications not cited by patent applications (Type C)



Rathenau Instituut

Note: Concept clusters are labelled CC. Node size indicates count of publications and count of INPADOC families. Node shading indicates timing of IA's entrance into concept cluster, white=early, grey=middle, black=late. Edge thickness indicates number of citing INPADOC families in the INPADOC cluster. Edge colour indicates period of citation, grey=early, dashed=middle, black=late.

Figure 5 demonstrates more clearly IA's direct knowledge and skill contributions to the technologies, as is shown by citations to publications authored by IA in the concept clusters. As also found in Figure 4, the concept clusters are cited at different stages of development. In Figure 5, however, the concept clusters also contain NPLRs authored by IA. In many cases, concept clusters containing NPLRs authored by IA are cited during the early stages of the span of a concept cluster, and others in the middle stages. Many of these come from the same research streams (for example CC d/4, d/1 & d/6 are all from research stream **Rd**) and are cited by all three technology groups. In the case of CC d/4, there are some transitive similarities to d/1 and d/6, and IA only begins publishing in the middle stages of that concept cluster's lifespan.

This once again demonstrates the absorptive capacity of IA because the research by IA that is necessary to the technologies is often cited early. Some of his research is cited later on. These publications appear in the middle stages of the concept clusters, but are eventually cited. In other words, aspects of his overall research have been necessary for the technologies and in areas in which he was not cited and/or active, he began research that eventually led to it being incorporated and cited.

Summarising the results seen in Figures 3-5, the scientific publications cited by the patent applications in the INPADOC clusters stem from three sources. These sources include 1) publications cited by the patent applications but not authored by IA, 2) publications authored by IA but not cited by the patent applications, and 3) publications by IA that are cited by the patent applications. The composition of the concept clusters and the period of citing by the INPADOC clusters indicate the relevance of the concepts to the technologies at different times. The entry of IA publications into these concept clusters indicates IA having a degree of similarity in knowledge and skill sets to the cited publications. The period of entry by IA's publications indicate the adoption of these skills and knowledge by IA. As per the definition by Cohen & Levinthal (1990), absorptive capacity is the ability "… to recognize the value of new, external information, assimilate it, and apply it to commercial ends [*and this*] is critical to its innovative capabilities." In this sense, the entry by IA publications into the NPLRs at varying time periods is indicative of the ability of IA to recognise, assimilate and integrate external knowledge into his own skill sets and knowledge base, and these integrations consequently become visible in the scientific background of the patent applications. We also observed IA starting new research in order to improve his absorptive capacity where he seemed to have gaps in knowledge and skills.

3.6 Summary and conclusion

The diverse characteristics of knowledge production, incorporation and dissemination relating to product development result in a complex model of knowledge capture. Previous methods used to investigate knowledge transfer have focused on the facilitating conditions with little concern paid to whether there is any actual knowledge transfer. In this paper, we have explained and developed a method to demarcate and track knowledge transfer. We have done so by combining and modifying existing techniques and supplementing them with new methodological tools. The resulting method allows us to address the more complex aspects of knowledge capture mechanisms - as illustrated with a stylised start-up or spin-off case.

With our methods for data processing, clustering and visualisation, we can demonstrate the thematic and theoretical links of the inventor's patent output and the inventor's knowledge base and skill sets, as represented by their publication output. This marks a departure from the problems of previous methodologies that relied on individual-specific direct citations from patent applications to literature in order to determine the theoretical influences of an individual or a field in general (Meyer, 2002).

Our method allows for a close examination of the multidirectional aspects of linkages between science and technology. This provides a quantitative measure of how effectively, and from where exactly, an idea generated in academia makes its way into an industrial application or, conversely, how skills and knowledge developed in application may be followed by new lines of research generating new scientific knowledge and skills.

Our approach to determining the absorptive capacity of an individual allows us to evaluate the utilisation of scientific knowledge by individuals and their eventual application in technology. The methodology accounts for the influence of co-inventors on the combination of knowledge required for technological output. This allows us to determine the degree and field of contribution from the respective inventors in terms of the base knowledge required for the development of a technology.

We explained and demonstrated the methodology using a stylised case study in which one individual is responsible for much of the spin-off firm's growth and success and bridges both the academic and industrial aspects of knowledge transfer. His research is both fundamental (at his university setting) and applied (in the firms' appropriation and implementation). The researcher-inventor bridges the research environments, facilitating knowledge transfer and skills development between them. In further research we aim to investigate the same processes if there are more star or bridge scientists in one firm, and the effect of their overall contributions.

Our method is not without its shortcomings. Using in-text citations of patent documents requires a significant amount of cleaning due to the differing citation reporting behaviours and requirements across patenting offices. The correct assignment of authors and inventors to publications and patents requires a significant amount of disambiguation. Initially this was done by algorithmic means (Gurney et al., 2012) and then checked manually, which was a time-consuming process. The inclusion of all the NPLRs in the patent applications introduces a level of uncertainty because some of the NPLRs are not directly related to the technologies. Many NPLRs are very general in nature and address only the fundamental background of the technologies. These NPLRs are difficult to identify without comprehensive expert examination of the patents but are still included.

This new method of mapping science and technology output and the relationships between them deepens our understanding of the level of contributions made by individuals and firms, and also by specific institutional policies and models. If a firm or an individual carries out research within a specific research climate or environment, by utilising this methodology one would expect to see the overall publishing and patenting activity, and the links between the two, to vary according to the research climate or environment. This therefore enables empirical investigation of the influence of the environment on knowledge transfer and absorptive capacity.

3.7 References

Arundel, A. (2001). The relative effectiveness of patents and secrecy for appropriation. *Research policy, 30*(4), 611-624.

Arundel, A. & Kabla, I. (1998). What percentage of innovations are patented? Empirical estimates for European firms. *Research policy, 27*(2), 127-141.

Audretsch, D.B. et al. (2005). University spillovers and new firm location. *Research Policy, 34*(7), 1113-1122.

Baba, Y. et al. (2009). How do collaborations with universities affect firms' innovative performance? The role of "Pasteur scientists" in the advanced materials field. *Research Policy, 38*(5), 756-764.

Bastian, M. et al. (2009). *Gephi: An open source software for exploring and manipulating networks*. Paper presented at the International AAAI Conference on Weblogs and Social Media.

Blondel, V.D. et al. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*, P10008.

Bozeman, B. (2000). Technology transfer and public policy: a review of research and theory. *Research Policy, 29*(4-5), 627-655.

Braam, R.R. et al. (1991). Mapping of science by combined co-citation and word analysis. I. Structural aspects *Journal of the American Society for Information Science and Technology 42*(4), 233-251.

Branstetter, L. (2005). Exploring the link between academic science and industrial innovation. *Annales d'Economie et de Statistique*, 119-142.

Brouwer, E. & Kleinknecht, A. (1999). Innovative output, and a firm's propensity to patent.: An exploration of CIS micro data. *Research policy, 28*(6), 615-624.

Callaert, J. et al. (2006). Traces of prior art: An analysis of non-patent references found in patent documents. *Scientometrics, 69*(1), 3-20.

Callon, M. et al. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics, 22*(1), 155-205.

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

Courtial, J.P. et al. (1993). The use of patent titles for identifying the topics of invention and forecasting trends. *Scientometrics 26*(2), 231-242..

Engelsman, E.C. & van Raan, A. F. J. (1994). A patent-based cartography of technology. *Research policy, 23*(1), 1-26.

Fleming, L. & Sorenson, O. (2001). Technology as a complex adaptive system: evidence from patent data. *Research Policy, 30*(7), 1019-1039.

Garfield, E. & Welljams-Dorof, A. (1992). Citation data: their use as quantitative indicators for science and technology evaluation and policy-making. *Science and Public Policy, 19*, 321-321.

Geuna, A. & Muscio, A. (2009). The governance of university knowledge transfer: A critical review of the literature. *Minerva, 47*(1), 93-114.

Gurney, T. et al. (2012). Author disambiguation using multi-aspect similarity indicators. *Scientometrics, 91*(2), 435-449.

Gurney, T. et al. (2013). *Knowledge capture mechanisms in bioventure corporations: a case study.* Paper presented at the 14th International Society of Scientometrics and Informetrics Conference 2013, Vienna.

Hicks, D. (1995). Published papers, tacit competencies and corporate management of the public/private character of knowledge. *Industrial and corporate change, 4*(2), 401-424.

Horlings, E. & Gurney, T. (2012). Search strategies along the academic lifecycle. *Scientometrics*, 1-24.

Karki, M. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information, 19*(4), 269-272.

Lancichinetti, A. & Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical Review E, 80*(5), 056117.

Langrish, J. et al. (1972). *Wealth from knowledge: a study of innovation in industry*: Halstead Press Division, Wiley.

Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy, 18*(4), 209-223.

Leydesdorff, L. (2008). Patent classifications as indicators of intellectual organization. *Journal of the American Society for Information Science and Technology, 59*(10), 1582-1597.

Magerman, T. et al. (2010). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics, 82*.

Martinez, C. (2010). *Insight into different types of patent families*: OECD.

Merton, R. K. (1957). Priorities in scientific discovery: a chapter in the sociology of science. *American Sociological Review, 22*(6), 635-659.

Meyer, M. (2002). Tracing knowledge flows in innovation systems. *Scientometrics, 54*(2), 193-212.

Meyer, M.S. (2001). Patent citation analysis in a novel field of technology: An exploration of nano-science and nano-technology. *Scientometrics, 51*(1), 163-183.

Mogoutov, A. et al. (2008). Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics, 2*(4), 341-353.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization science, 5*(1), 14-37.

Palla, G. et al. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*(7043), 814-818.

Pavitt, K. (1988). Uses and abuses of patent statistics. *Handbook of quantitative studies of science and technology*, 509-535.

Ponomariov, B. & Boardman, C. (2012). *Organizational behavior and human resources management for public to private knowledge transfer: an analytic review of the literature*: OECD Publishing.

Rosenberg, N. (1990). Why do firms do basic research (with their own money)? *Research Policy, 19*(2), 165-174.

Schmoch, U. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators. *Scientometrics, 26*(1), 193-211.

Schmookler, J. (1966). *Invention and economic growth*: Harvard University Press Cambridge, MA.

Somers, A. et al. (2009). *Science Assessment Integrated Network Toolkit (SAINT): A scientometric toolbox for analyzing knowledge dynamics*. The Hague: Rathenau Instituut.

Tamada, S. et al. (2006). Significant difference of dependence upon scientific knowledge among different technologies. *Scientometrics, 68*(2), 289-302.

Tijssen, R.J.W. (2001). Global and domestic utilization of industrial relevant science: patent citation analysis of science-technology interactions and knowledge flows. *Research Policy, 30*(1), 35-54.

Tijssen, R.J.W. (2006). Universities and industrially relevant science: Towards measurement models and indicators of entrepreneurial orientation. *Research Policy, 35*(10), 1569-1585.

Tijssen, R.J.W. & Van Raan, A.F.J. (1994). Mapping changes in science and technology. *Evaluation Review, 18*(1), 98-115.

van den Besselaar, P. & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: a method and an exploratory case study. *Scientometrics, 68*(3).

White, H.D. & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science (1986-1998), 49*(4),327-355.

Whittaker, J. (1989). Creativity and conformity in science: titles, keywords and co-word analysis. *Social Studies of Science, 19*(3), 473.

4 Knowledge Network Influences on the Development of Drug Discovery Technologies: A Longitudinal Case Study[1]

Abstract

Mechanisms of knowledge transfer from academia to industry have long been debated. The knowledge inputs required may stem from research conducted many years prior to a technology being adopted and adapted by industry. A supporting knowledge base is required to facilitate knowledge transfer. In this paper we utilise the publishing and patenting history of an individual scientist to demonstrate in detail how new technologies emerge from the work of academic inventors and from their cognitive environment. It provides a detailed description of the knowledge within these technologies. In particular, we will address the role of absorptive capacity in priming their development. We find clear linkages between the technologies in their present form and the long-past specific outputs authored by the individual, and that the knowledge contained therein has undergone varied degrees of transformation. The individual scientist demonstrates a high level of absorptive capacity, incorporating exogenous knowledge into their own knowledge base.

4.1 Introduction

In Nelson's seminal essay entitled "The market economy, and the scientific commons" (R. Nelson, 2004), technologies need to be understood as: *"[I]nvolving both a body of practice, manifest in the artefacts and techniques that are produced and used, and a body of understanding, which supports, surrounds, and rationalizes the former." (p.457)*

In knowledge and technology transfer, analyses address a multitude of aspects and levels of science-technology interaction (Bozeman, 2000; Ponomariov & Boardman, 2012). A minute analysis of knowledge transfer mechanisms and mediums - such as using tacit and codified knowledge, R&D networks, formal and informal collaborations - runs into many difficulties. These difficulties stem from the enormous complexity of the knowledge that is transferred. In most cases, the final technological product is the result of heterogeneous knowledge inputs and its accretion over time into a coherent system (Nelson, 2004).

In this paper, we focus on the knowledge inputs of a firm and on the nature of the quanta of knowledge and information themselves. We will add to the current literature on knowledge transfer by examining in detail the specific knowledge and technologies involved, by means of an analysis of the actual uptake and implementation of the associated knowledge concepts into the technologies developed.

---

1 This chapter has been published as Gurney, T. et al. (2013). *Knowledge capture mechanisms in bioventure corporations: a case study.* Proceedings of 14th International Society of Scientometrics and Informetrics Conference, Vienna. It has been submitted for publication in Research Policy.

We have developed a method to discern the knowledge contributions of inventors and scientists to a corpus of patents and the technologies they represent (Chapter 3). The method is applied to the two main output indicators typically used in other studies, those of patents and publications. Publications are the typical output of scientific endeavours, and patents are the technological result of the application of the results of those endeavours. The concepts and practices embodied and codified in the publications and patents were linked to each other through the citations to literature found in the patent documents. By linking the two corpora of scientific and technological knowledge, we are able to address our research question: *To what degree does an existing knowledge base contribute to the development of novel technologies and how can we effectively measure these contributions?*

As such, we aim to demonstrate the origins of the knowledge contributions to the development of an idea over time, from its inception, through its transformation and finally to its application in a technological product. The initial sections of this paper discuss the multiple aspects of absorptive capacity, knowledge transfer and transformation, including how scientific knowledge is incorporated into practices, skill sets and artefacts. We then discuss the national policy context of our test case study, and provide a brief history of our test case study. Following this, we briefly summarise our previous methodology, along with descriptions of the indicators we use, followed by the visualisation and clustering techniques employed in our analysis. Our results and conclusions follow, ending with our discussion and implications for further analyses and policy.

4.2 Conceptual Framework

The most common and widely cited knowledge transfers, capture mechanisms and inputs are patents, publications, informal and formal interactions, personnel hiring, licensing, R&D collaborations, contract R&D and consulting (Cohen et al., 2002). In each of these mechanisms, the medium of knowledge transfer is either codified (such as, for example, patents and publications) or tacit (such as, for example, R&D collaborations and personnel hiring). A third medium, that of embedded knowledge, resides in the material aspects, such as new equipment (Gorman, 2002). Key to the reception and implementation of these mediums is the absorptive capacity of the unit under study.

4.2.1 Absorptive capacity

The organisational infrastructure required to facilitate the development, transfer and capture of knowledge depends heavily on its recipient. The recipient is understood to demonstrate a need for 'absorptive capacity' (Cohen & Levinthal, 1990) which is described as "[t]he ability of a firm to recognize the value of new, external information, assimilate it, and apply it to commercial ends [*and this*] is critical to its innovative capabilities," (p.128).

There are two key aspects involved in absorptive capacity, firstly on an individual level, in that the absorptive capacity of a firm is tied to its constituent individuals' absorptive capacity, i.e. the right personnel are in place to take advantage of incoming information. In this aspect, the communication infrastructure between a firm and its external environment is key, with select individuals acting as gatekeepers, much like the star scientists of Zucker and Darby (1996) or the core scientists of Furukawa and Goto (2006). In the case of a dedicated university-industry spanning role, as might be fulfilled by a bridge scientist (a scientist who is active within both academia and

industry), the absorptive capacity could be bolstered by a more active search process. As Cohen and Levinthal (1990) state, "Beyond diverse knowledge structures, the sort of knowledge that individuals should possess to enhance organizational absorptive capacity is also important. Critical knowledge does not simply include substantive, technical knowledge; it also includes awareness of where useful complementary expertise resides within and outside the organization" (p.133).

Secondly, on an organisational level, the requirements for absorptive capacity are related to the network-wide communication structure, such as "the relationships between corporate and divisional R&D labs or, more generally, the relationships among the R&D, design, manufacturing, and marketing functions" (Cohen & Levinthal, 1990) (p.134). In essence, this form of absorptive capacity links the interactive, exchange aspects of innovation to the communication aspects. This form relies on the positive and/or negative feedback between departmental/divisional entities to further develop an idea or product. In this paper, we do not focus on this level of absorptive capacity - which we do elsewhere (Chapter 3 and (Lanciano-Morandat et al., 2009)) - but on the knowledge capture mechanisms carried out by the star and bridge scientists discussed above.

The concept of absorptive capacity was expanded on significantly by Zahra & George (2002) who include potential and realised absorptive capacity. This expansion has clarified what absorptive capacity encompasses, by defining its various dimensions. We utilise Zahra & George's dimensions in our case study. These dimensions include, for potential absorptive capacity: *Acquisition* - details the role of prior knowledge or capabilities and the infrastructure already in place; and *Assimilation* - exogenously generated knowledge (by others outside the firm) needs to be understood prior to incorporation within a firm. For realised absorptive capacity: *Transformation* - the ability to combine knowledge generated exogenously and endogenously (within the firm) to create novel fundamental or applied knowledge; and *Exploitation* - the usage of novel knowledge generated during transformation. The outputs of exploitation can be patents, publications, products or new processes.

4.2.2 Proxies of knowledge input and output

All four dimensions of absorptive capacity introduced by Zahra & George (2002) include patents and publications as typical knowledge mediums (as well as outputs). We use these codified knowledge mediums as proxy indicators of exogenous and endogenous knowledge. These proxies embody and detail the knowledge units acquired, assimilated, transformed and finally exploited. The use of patents and publications as such proxies is well developed and details of which are given below.

The use of patents as indicators was pioneered by Schmookler (1966) with many applications following (e.g. Griliches (1998), Schmoch (1993) and Fleming (2001)). Patents have been used as indicators for multiple purposes as they are considered detailed evidence of technological progress (Tijssen, 2002). They are inherently scalable (Narin, 1995), contain multiple metadata useful for analysis (Nelson, A.J., 2009), and are commonly regarded as substantive links to R&D activity (Ahuja & Katila, 2001). There are drawbacks in that not all innovations are patented (Arundel, 2001; Arundel & Kabla, 1998; Pavitt, 1988). However, they arguably remain the best indicators of R&D input (Narin, 1994).

When patents are used as indicators, the analysis is typically based on the metadata found in patents. Title words, abstract words and keywords (Courtial et al., 1993; Engelsman & van Raan, 1994), patent classifications (Leydesdorff, 2008; Tijssen & Van Raan, 1994), and patent/non-patent citations (Karki, 1997; Meyer, M.S., 2001) have all been used extensively. We use the PatSTAT database (April 2011 version), prepared by the EPO, because it is widely available, has global coverage and is data-rich and comprehensive.

Publications serve as the primary indicators for the defining characteristics and development of science. They are the most visible outcome of scientific endeavours, and an extensive range of indicators and methodologies have been developed. The analysis of publications shares a number of analytical approaches with patent analyses, such as word mapping (Callon et al., 1991) and citation analysis (Garfield & Welljams-Dorof, 1992; White & McCain, 1998). Sequential usage (Braam et al., 1991) and combinations (van den Besselaar & Heimeriks, 2006) of title words and cited references of publications has also become a common analytical technique.

4.2.3 Non-patent literature references

As part of the patenting process, listing prior art is a requirement across all patenting offices (although the completeness of submitted prior art varies between offices (Criscuolo & Verspagen, 2008)). Citation studies using patent-to-literature citations have progressed immensely (Meyer, M., 2000, 2002; Meyer, M.S., 2001; Narin, 1976, 1994). These studies examined in detail the literature citing proclivity and distribution across fields. Key to these studies was the recognition that references to literature in patents do provide a substantive link between the technologies and the sciences.

Non-patent literature references (NPLRs) exhibit different characteristics based on their source and who includes the reference. The scientific nature of NPLRs has been examined by Callaert et al. (2006). They found that there may be occurrences of citations to non-scientific literature but, overall, most citations are to peer-reviewed journals. NPLRs may come from applicants or examiners and have typically been treated as being of differing importance (Karki, 1997). The complex negotiation processes between applying for a patent and the granting of that patent result in the strategic inclusion (or exclusion) of specific references by the applicant. Examiners must rely on references supplied by the applicant and their own examination process to ensure a complete check. Following this, studies have typically chosen the front-page references (on the patent document), i.e. the examiner references, over the applicant references. In our study we choose to utilise both sources of references, bibliographic-NPLRs (examiner) and in-text NPLRs (applicant). We do so as it is generally understood and accepted that the presence of citations to literature in patent documents indicates a cognitive link to, or awareness of, the related scientific concepts (Tijssen, 2001), regardless of the source of the NPLRs. By combining the two sources of NPLRs we aim to provide a more comprehensive view and call upon the judgement of not only the inventor but also the examiner as to what is relevant.

4.2.4 Publishing and patenting in academia and industry

In studies such as ours, which examines patents or publications across the academic-industrial divide, it is important to note that each sphere maintains different approaches to each aspect of knowledge production. When comparing university-based and firm-based scientists and their

propensity for the different types of knowledge production, it is important to consider the underlying motivations. University-based scientists publish primarily to extend their professional and intellectual prowess (through which resources are allocated for future projects) and regular publishing is considered a requirement. With patenting, there has been a recent explosion of sorts in the rate of university patenting. This has been linked to institutional and national level changes (Owen-Smith & Powell, 2003; Zucker, L.G. & Darby, 1996), and the increased interest in spin-offs and IP spin-outs generated in academia (Owen-Smith & Powell, 2003; Zucker, L.G. & Darby, 1996; Zucker, L.G. et al., 1999).

Firms benefit from carrying out their own basic research, as they become intimately involved in the fundamental aspects of their own applied research (Rosenberg, 1990) and the output of this basic research typically results in publications. With firm-based publishing efforts, the underlying motivation is generally a directed and concerted effort from within the corporate infrastructure, as the firm stands to gain (or lose) more from the publication process than the author. For example, higher rates of approval of patents (McMillan et al., 2003), a window and source into various fields (Schartinger et al., 2002) and stronger ties with future progenitors of knowledge (Zucker, L.G. & Darby, 1996). Publishing in firms has also been shown to be strongly linked to future patenting areas and recruitment efforts (Hicks, 1995).

Industry has particularly realised the benefits of collaborative efforts with academia, as multiple studies have demonstrated positive results in terms of innovation output, ranging from the life sciences sector to the nanosciences (Baba et al., 2009; Meyer, M., 2007).

4.3 Knowledge utilisation

In this case study we aim to comprehensively describe the knowledge utilisation processes, including absorptive capacity, transfer and capture within a company, using a methodology developed previously in Chapter 3 of this thesis. We use this methodology to investigate the research question: *To what degree does an existing knowledge base contribute to the development of novel technologies and how can we effectively measure these contributions?*

We address this question by developing descriptors (indicators) of the degree of similarity between a researchers' publication corpus and the patent applications of which he is an inventor. These descriptors, in general terms, describe the level of endogenous versus exogenous (to the individual or research group) knowledge influences on the patent output. More precisely, they relate to the receptivity of the researcher/inventor and the firm involved to the sources of knowledge, and the specific knowledge that was transferred or captured. These include (i) the researcher/inventor's own knowledge that he or she brings in; (ii) knowledge brought in through research collaboration with others; (iii) knowledge produced by others, but which is similar enough to the companies' knowledge base to absorb it directly; and (iv) new research conducted to be able to absorb knowledge which is rather dissimilar to the existing knowledge base. The descriptors are:

1. The reputational and applicability aspects of the scientific base work (Hullmann & Meyer, 2003) conducted by an individual;
   - The reputational aspect (as defined by the number of citations) and applicability aspect (as

defined by the proportion of the researcher/inventor's work cited by the patent applications versus their total corpus) determine the quality and relative importance of the researcher/inventor's research to the technology.

2. If the individual's overall research trajectory is, or is not, located in the field(s) of research necessary for the technologies;
   - This provides detail of the knowledge base cited by the patent applications and highlights the similarities (if any) between the publications cited by the patent applications and the overall research corpus of the researcher/inventor.

3. The markers for the other fields of science that are being utilised by the technologies (Karki, 1997; Schmoch, 1993);
   - This describes knowledge sourced from outside the researcher/inventor's own expertise or network. This is defined by what research is cited by the patent applications and does not fall into the fields of research of the researcher/inventor.

4. The level of input by collaborators;
   - Externally sourced knowledge, from collaborators or co-inventors, may be required for the technologies. These contributions may be identified by the cited literature that is authored by collaborators/co-inventors without the researcher/inventor as an author.

5. The degree that the knowledge features (such as concepts, knowledge bases and, to a certain extent, skill sets) utilised by the technologies are shared amongst their sources (the inventor, his co-inventors, or other researchers);
   - The similarities or differences between the knowledge cited by the patent applications and the overall corpus of the researcher/inventor may indicate whether the required research and the associated skill sets are already present or need to be developed.

6. Whether the individual incorporated skill sets that were acquired during the development of the technologies and applied them to further his or her fundamental scientific research by knowledge creation feedback (Fischer, 2001; Tijssen, 1998).
   - If research conducted by the researcher/inventor after applying for a patent displays a degree of similarity to the publications cited by the patent applications - in which the researcher/inventor had no previous presence - this indicates that the knowledge and research skills obtained during the development of the technologies were applied in his or her further work.

We have chosen a case study in which an individual occupies a significant bridging role between academia and industry. This role allows us to effectively isolate his contributions to both the technologies and the underlying sciences involved, and allow us to fully investigate the descriptors mentioned above. The method can also be applied on more hybrid cases, but for reasons of clarity we selected this one.

4.4 Case study selection and history of Japanese biotechnology

4.4.1 Case study selection

Our case study involves a prominent Japanese biotechnology researcher, Professor Yusuke Nakamura, who is heavily involved in cancer therapeutics at the University of Tokyo, where he was head of the Human Genome Center. Nakamura founded OncoTherapy Science Inc. (OTS) in April 2001 to "[…] contribute in research and development of anti-cancer medicine, cancer therapy and cancer diagnosis based on oncogenes and proteins".[2] OTS's business outline is: *"…to provide innovative anti-cancer medicines with higher efficacy and a minimum risk of adverse events based on the comprehensive research on cancer genomics and biomedical analysis conducted by Nakamura of Human Genome Center, Institute of Medical Science, the University of Tokyo."*[3]

Although Nakamura is not currently listed on the board of directors, he maintains direct links between his research at the University of Tokyo and research conducted at OTS. This direct link between academia and industry as manifested by Nakamura is the primary reason for choosing Nakamura and OTS. It enables us to draw upon his extensive publishing history as well as his numerous patenting activities, both at the University of Tokyo and OTS.

Below we provide some background information on the Japanese biotechnology sector from our own research into public documents to provide more detail on our selection of this case study.

4.4.2 Japanese firms in biotechnology research

Japan has a long tradition of biotechnology and can be regarded as a biotechnology-orientated country. In 1908, Dr Kikunae Ikeda at the University of Tokyo found that monosodium-glutamate is the true substance of the "Umami" taste. In the same period, Mr Saburosuke Suzuki was extracting iodine from konbu (seaweed), for pharmaceuticals. Collaboration between Dr Ikeda and Mr Suzuki led to the foundation of Ajinomoto, one of Japan's first venture companies deriving from cooperation between industry and academia. This example illustrates two important aspects: that science was driven to application and that Japan was amongst the pioneers of biotechnology-based technology transfer in the early 20th century. Many such firms were developed (and remain very active in all fields) such as Takeda Pharmaceutical Company Limited, Kyowa Hakko Kirin Co., Ltd., Astellas Pharma Inc. and Daiichi Sankyo Co., Ltd. to name a few.

During the 1980s, Japan initiated and developed international collaborative programmes to improve its life sciences efforts in pre-competitive research. During the period in office of Prime Minister Yasuhiro Nakasone (1982-1987), the Human Frontier Science Program was started (1987), the Japan Key Technology Center was established (1985), the Protein Engineering Research Institute was established (1986), amongst others. As a result, the life sciences in Japan became more internationalised with an increase of foreign scientists at different levels, from post-docs to senior scientists. However, this strategy did not have the desired increased impact on promoting

2 http://www.oncotherapy.co.jp/eng/corporate/enkaku.html
3 http://www.oncotherapy.co.jp/eng/corporate/business.html

biotechnology-based knowledge transfer. Japan still lagged behind the USA and Europe in this aspect. During this period of transformation, there were few efforts by researchers to obtain patents. Sometimes inventions developed at universities were transferred free of charge from faculty members to companies with which the faculty members have close relations. These practices are a far cry from transferring technologies developed at universities to the most appropriate corporations. There were no systematic technology transfer policies in place from academia to industry, nor biotech start-ups developed from academia.

4.4.3 Japanese innovation policies

Following the collapse of the Japanese economy in the early 1990s, in order to create new industries for economic growth, the decision was taken to appropriate funds to support research and development at universities, starting in 1995 The Science and Technology Basic Law was enacted in November 1995, and the Cabinet approved the first phase of the Science and Technology Basic Plan in July 1996. Universities were increasingly expected to serve as the source of technological innovation in order to revitalise the economy, and society increasingly needed to see returns on research expenditure at universities, which grew amid the economic slump. Under these circumstances, measures related to industry-university cooperation were promoted in the late 1990s. The concept of technology-licensing organisations (TLOs) emerged in 1998 as the symbol for providing patents on inventions developed at universities. The law on promotion of transfer of technology-related research results from universities and other institutions to private corporations (the law for technology transfer from universities) was enacted in 1998 and stipulates conditions for TLOs to receive government designation. Activities related to industry-university cooperation expanded following the passage of this law.

In 2000, the rules and regulations of the National Personnel Authority were revised to increase willingness to set up start-ups led by academics. This enabled faculty members at national universities to also serve as corporate executives with the aim of commercialising research results. In 2002, under Prime Minister Junichiro Koizumi, the Intellectual Property Policy Outline was established, and the Basic Law on Intellectual Property was enacted. On 8 July 2003, the government's Intellectual Property Policy Headquarters announced the Strategic Program for the Creation, Protection and Exploitation of Intellectual Property. On 15 July, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) announced 34 institutions that qualified for the University Intellectual Property Headquarters Development Programme. These changes showed that universities had entered a new era in terms of intellectual property rights.

Along with this trend of emphasising the importance of intellectual property, a 'Japanese cluster policy,' aimed at promoting innovation in a specific region by linking technology seeds in university or public research institutions with corporations, was initiated by two ministries during the 2000s. The Ministry of Economy, Trade and Industry (METI[4]) started the Industrial Cluster Initiative in 2001 and MEXT[5] started the Intellectual Cluster Initiative in 2002. An example of this

4 Following the Central Government Reform in January 2001, the Ministry of International Trade and Industry (MITI) was renamed METI.

5 Owing to the Central Government Reform in January 2001, the Ministry of Education, Science and Culture was merged with the Science and Technology Agency to establish MEXT.

was the Kansai region, where one could observe a rapid accumulation of laboratories of large bio-related companies, bio start-ups, public research institutions and universities.

According to the 2008 report of the Japan Bio-Industry Association, there were 577 bio-venture companies in Japan as of 2007, and the landscape was heterogeneous, incorporating medical research, research support, consulting, environment, agriculture, and production of bio-molecules.

4.5 Method

To summarise the methodology developed in Chapter 3, the thematic and knowledge-based aspects of the patent application and publication data are linked to each other through similarities between the bibliographic and in-text NPLRs of the patent applications and the publication corpus of the inventor. The necessary data and processes will be further explained in the following sections.

4.5.1 Data collection

For our publication and patent data, we use the European Patent Office (EPO) patent database PatSTAT (September 2011 version) and Thomson Reuters' (ISI) Web of Science (WoS) publication database.

The sources and types of data come from:

1. Patents - we extracted all patent applications with OncoTherapy listed as an applicant from the EPO PatSTAT database (2000-2008)[6] of all inventors;

2. Publications[7] - we downloaded all publications with OncoTherapy listed as an institution from WoS (all entries up to 2011); and all publications with Nakamura listed as one of the authors.

These base data were parsed using SAINT[8](2009) and managed in a relational database. We collected further data from the patents - specifically (where found):

3. In-text non-patent literature references (IT-NPLRs) - citations to publications within the body of the patent, but not always in the front-page reference list. These IT-NPLRs were automatically extracted from the full-text versions of the patent documents by custom software.

4. Bibliographic NPLRs (B-NPLRs) - these are citations included primarily by the examiner and added as front-page references.

We grouped the collected patent documents by INPADOC family[9]- and aggregated the data

---

6 We chose patents up to 2008 as there is considered to be a delay in the completeness of patent data in PatSTAT. 2008 was chosen as the last year as we could be more certain that it included all possible patent data.

7 English language only

8 SAINT (Science-system Assessment Integrated Network Toolkit - a Rathenau Instituut open-source software suite designed to parse, clean and organise bibliometric data to be later used in relational database software such as MS Access and MySQL.

9 INPADOC extended families are grouped by Paris Convention priorities, domestic continuations and technical relations. The INPADOC family serves to aggregate patents protecting the same or related inventions, represented by different applications over time or different patenting offices.

associated with each application to the parent INPADOC family. This was done to overcome the disparities and lack of data associated with patent documents from some patent offices. As such, we choose to view the collective patent documents and INPADOC families as representing a specific technology (Martinez, 2010). We extracted the patent application metadata up to December 2008 from the EPO's PatSTAT database.

For each of the patent documents extracted from PatSTAT, we extracted the non-patent literature references (NPLRs) using custom software. The software identified and downloaded the full-text versions from the EPO web portal, and parsed and extracted both the in-text NPLRs (IT-NPLRs) and bibliographic NPLRs (B-NPLRs).

As far as possible, we located and downloaded the ISI WoS publication equivalents of all the NPLRs and added them to the existing publication data set. Some NPLRs could not be linked to WoS publications as they had insufficient identifying data, such as the author name only, or author name and year only, or journal name and year only. However, these were minimal as most NPLRs contained enough data to accurately link them to publications found in WoS.

The parsed publication corpora were grouped into a single relational database, recording the origins of each document within the combined set.

4.5.2 Similarity calculations and clustering

Publications

The similarities between publications (both NPLR and Nakamura's) were calculated based on their shared cited reference and title word combinations using a method developed by Van den Besselaar and Heimeriks (2006). We constructed a network using the publications as nodes and the edges representing the degree of similarity as calculated above. The research streams of publications within the network were assigned using a community detection algorithm developed by Blondel et al. (2008). Once the initial research stream assignment was completed, the general streams were isolated and the community detection algorithm was run again to produce smaller concept clusters (Gurney et al., 2012).

INPADOC families

The INPADOC families were clustered using the International Patent Classifications (IPC) codes added by the examiner to the patent application at the time of application. The IPC classification codes are internally orientated search codes for examiners to assign patent applications to different classes. The use of IPC codes as tokens in similarity calculations to determine knowledge-relatedness is an extensively developed method (Breschi et al., 2003; Jaffe, 1986). We examined in detail the patent titles and claims associated with each patent application within each INPADOC family to determine the clinical application, general methodology and target disease. Then we recorded the resulting clusters of INPADOC families.

Linking patent families and publications

The NPLRs were co-located within the general research streams based on the level of similarity of shared title word and cited reference combinations. The shared knowledge features, such as

concepts, knowledge bases and, to a certain extent, skill sets involved can be elucidated through the degree of similarity between the publications. By linking the INPADOC families to the general publication communities in which their NPLRs are co-located, we can infer that there is at least a degree of shared knowledge features between the publication community and the citing INPADOC families. For more specific knowledge features, the second layer of concept clusters provided a finer-grained view into the communities.

The source composition of publications varies within each concept cluster. In our case study, in which Nakamura is the primary producer of the publications, each concept can potentially contain a mixture where varying proportions of source publications imply differing levels of imparted or similar knowledge features of the publications and the INPADOC families. These concept clusters include NPLRs where a) Nakamura is the author; b) Nakamura is not the author and/or Non-NPLRs authored by Nakamura but not cited by the patent document.

Visualisation technique

To visualise the research streams over time, we employ a method introduced by Horlings & Gurney (2012) where 'cognitive communities' or 'research trails' are isolated and mapped over time. This method of visualisation provides a clear view of what we call in this paper different 'research streams' and their mutual relations. This manner of visualisation also allows closer examination of an individual's contributions during various phases of their career trajectory, such as during their PhD, post-doc and professorial phases.

Using this method of analysis and visualisation, we are able to simultaneously place Nakamura's scientific output (publications) and technological output (patents) within the same frame. The linking method of clustering the NPLRs and researcher/inventor publication corpus allows us to comment directly on the similarity between the scientific work undertaken by Nakamura and the technological output of which he is a primary inventor.

4.6 Results

4.6.1 Patents and patent families

In total we collected 242 patent application documents via PatSTAT with Nakamura listed as inventor and OncoTherapy as assignee. The patent documents came from 90 INPADOC families, and were composed of 115 priority patents[10]. The priority patent applications were primarily filed in the USA (101 applications) and the rest in Japan (14 applications). The earliest patent filing date was March 2000, and the latest was November 2008. The maximum, minimum, average and median numbers of patent applications per INPADOC family were 23, 2, 5.3 and 4 respectively.

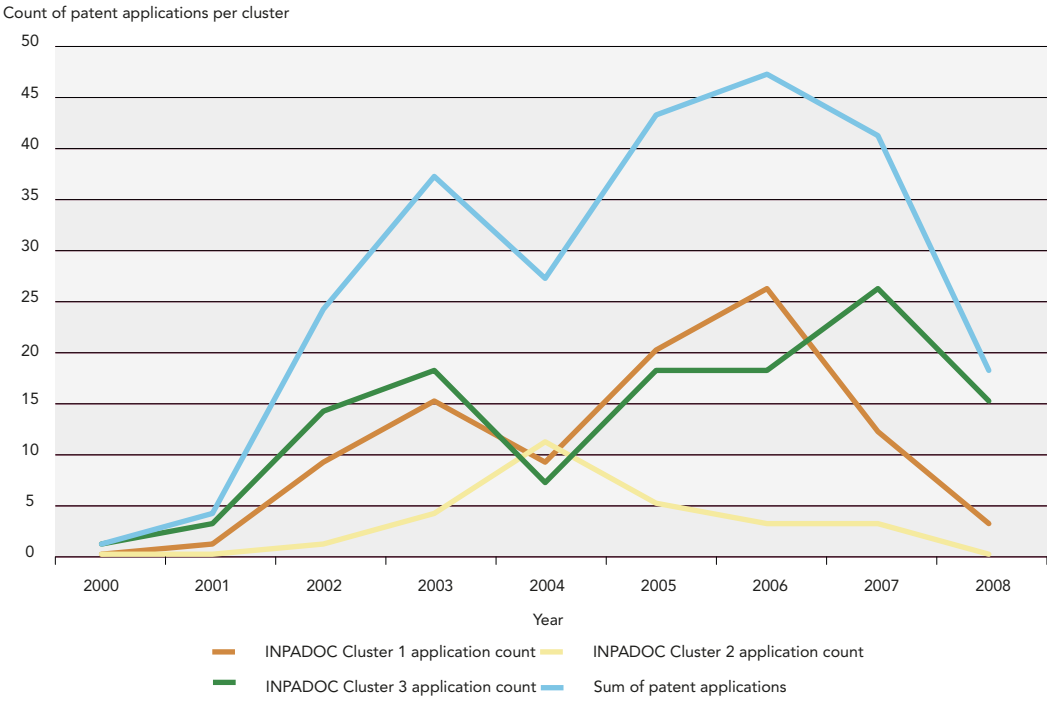
4.6.2 Clustering of INPADOC families by IPC

Three primary INPADOC clusters were found, using main group IPC data. The growth in the number of patent applications per INPADOC cluster is shown in Figure 1(a) and the count per year of unique families in each INPADOC cluster is shown in Figure 1(b). In Figure 1(a), the number of patent

10 Priority patents have been defined in this study as patent applications that have no earlier priority date.

applications in clusters 1 and 3 increased, whilst cluster 2 showed little increase. From 2003, this pattern reversed and the number of patent applications in cluster 2 eventually overtook clusters 1 and 3. This trend once again reversed from 2004 when cluster 1 became the dominant cluster until 2006 when cluster 3 overtook it. From Figure 1(b), in 2002 and 2004, the number of unique INPADOC families increased at a slower rate, suggesting a period of specialisation within OncoTherapy. Taken together, from 2004, the increased number of unique families and the increased number of patent applications suggest a period of diversification for clusters 1 and 3, whilst research in cluster 2 decreased overall. From 2006, there was less patenting overall, but a similar number of unique INPADOC families, suggesting that there was still diversity overall in the fields of technology being addressed, but no visible expansion in diversity.

Figure 1(a) INPADOC cluster patent application count.



Rathenau Instituut

Figure 1(b) INPADOC cluster family count.

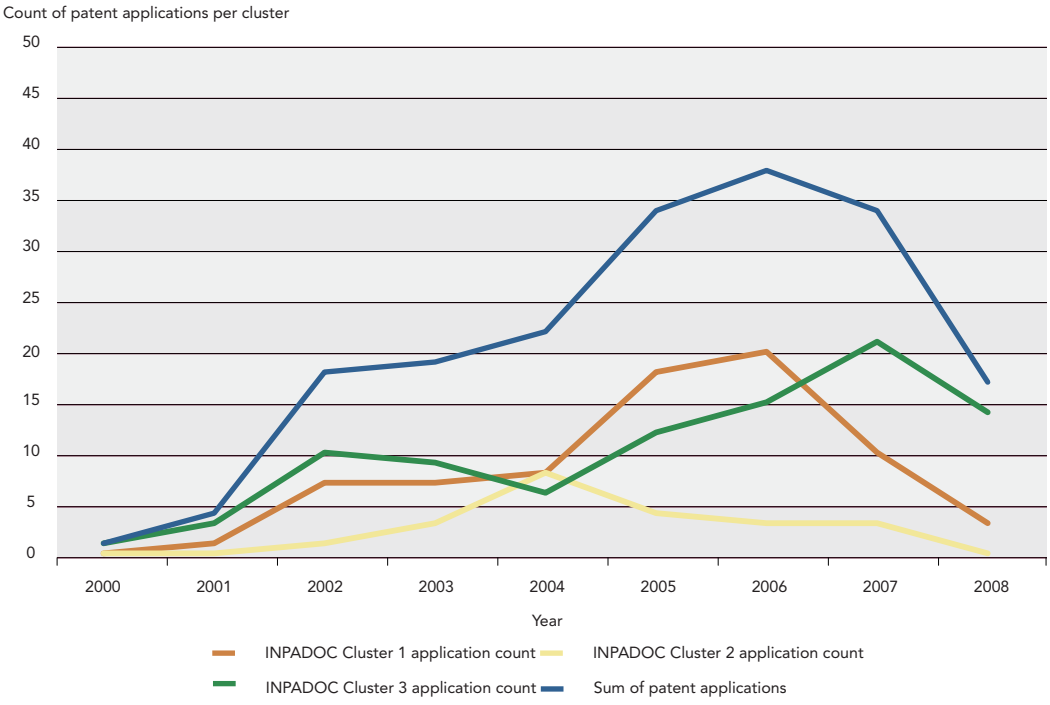
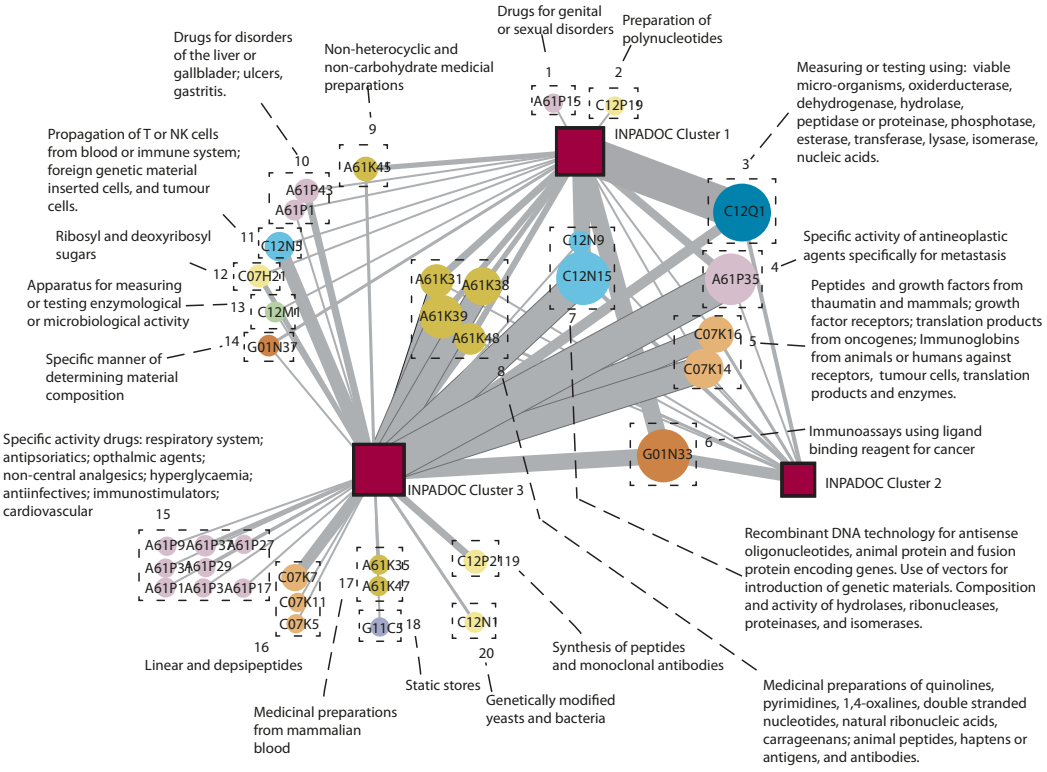


Figure 2 shows a 2-mode network of the aggregated INPADOC clusters and main group level IPC technological areas. Clusters 1, 2 and 3 contain technologies from eleven shared main group level IPC code areas, whilst clusters 1 and 3, and clusters 2 and 3, share seven and zero IPC code designations respectively. The 2-mode network in Figure 2 demonstrates the specific areas shared by each INPADOC cluster but also serves to highlight which clusters have specialised technological areas that are only applicable to each cluster. INPADOC cluster 2 does not address any unique areas, whereas both clusters 1 and 3 do.

Figure 2 Annotated 2-mode network of main group-level IPC and INPADOC family clusters



Rathenau Instituut

Note: Rectangle represents clusters, dashed rectangle represents main groups, circle represents families. Node size of INPADOC clusters indicates number of INPADOC families; size of main group IPC nodes indicates number of patent applications utilising the main group IPC code. Edge weight is the proportional count of number of patent applications utilising the main group IPC code. Annotations are summarised from the WIPO IPC classification descriptions.

As Figure 2 shows: at the main group IPC levels, the subject areas addressed by the INPADOC clusters relate primarily to the use of micro-organisms, enzymes, peptides and growth factors, recombinant DNA technologies and medicinal preparations using the peptides and RNA.

The overall description is in line with OncoTherapy's official stated research theme: the development of technologies related to gene expression analysis, identification of target genes, cancer peptide vaccines, antibodies for treatment and diagnosis, small molecule drugs and RNA medicines.

4.6.3 Publications and NPLRs

The methodological approach taken in this study considers, in tandem, the publications of Nakamura and the NPLRs found in the patents of OncoTherapy. The wider body of Nakamura's publications is discussed first, and followed by the NPLRs found both in the bibliographic (B-NPLR) section of the patent applications and the in-text references (IT-NPLRs).

Nakamura publications

Nakamura has published extensively with 931 publications over 33 years. His first publication was in 1977 and he published less than five publications per year until 1987. Between 1988 and 1994, he published between 5 and 10 publications a year and in his current phase, his publication count jumped to about 50 a year. His earliest phase of publishing coincided with his MD and PhD, and research fellowships. His middle phase was composed of an assistant Professorship at the University of Utah and becoming Head of Department at the Cancer Institute in Tokyo. His current phase coincides with his Professorship at the University of Tokyo and his directorships of both the RIKEN Center for Genomic Medicine and of the Human Genome Center at the University of Tokyo.

NPLRs

In total we were able to isolate 2037 NPLRs from the 242 patent applications. Of these 2037, there were 842 unique NPLRs. We were able to successfully match 525 of these NPLRs with the WoS. Of the matched NPLRs, there were 147 unique NPLRs found only in the bibliography, 313 unique NPLRs found only in the text and an overlap between the two of 65 NPLRs. Of the 525 matched NPLRs, 259 NPLRs are cited more than once, 73 are cited 5 or more times and 20 are cited 30 or more times. The most cited publication is cited by 41 different patent applications. The most cited publications come from the time period of 1996-2004 with less than 10% of NPLR citations going to publications older than 1996.

Publication clustering

We clustered the publications as described in the methods section. The resulting clusters represent research streams that differ in size and duration. The largest research stream (by count of Nakamura publications) is stream 2. Streams 1 and 2 contain both high numbers of Nakamura's publications and NPLRs. 55 (10.5%) of Nakamura's publications are cited as NPLRs - 19 in-text NPLRs, 18 bibliographic NPLRs, and 18 cited both in-text and bibliographic.

Streams that are almost exclusively NPLR-based include streams 4, 8, 9, 10 and 12. Nakamura's publications are the only publications in streams 15, 16 and 18. There are a few streams composed of only two publications, and these are 3, 6, 17 and 19. Stream 3's two NPLR publications are extremely highly cited (approximately 40 000 citations each). Table 1 provides a summary of the streams including the relative presence of NPLRs in each stream.

Table 1 Summary stream summary

Research Stream	Count Total (Nakamura/NPLR/Both)	Average Citations per Nakamura publication	Start	End	Trail Length (years)
1	157 (84/73/0)	25	1978	2011	34
2	273 (182/90/1)	28	1978	2007	30
3	2 (0/2/0)	-	1979	1987	9
4	85 (5/80/0)	20	1987	2008	22
5	169 (133/35/1)	51	1987	2009	23
6	2 (2/0/0)	12	1988	1989	2
7	135 (97/18/20)	15	1988	2011	24
8	15 (0/15/0)	-	1988	2005	18
9	78 (6/72/0)	23	1989	2007	19
10	8 (0/8/0)	-	1991	2002	12
11	6 (5/1/0)	24	1992	1995	4
12	15 (0/15/0)	0	1993	2005	13
13	159 (110/16/33)	50	1994	2010	17
14	8 (6/2/0)	32	1996	2005	10
15	15 (15/0/0)	35	1996	2005	10
16	20 (20/0/0)	47	1997	2003	7
17	2 (0/2/0)	-	1998	2001	4
18	183 (183/0/0)	55	1999	2011	13
19	2 (0/2/0)	-	1999	2000	2

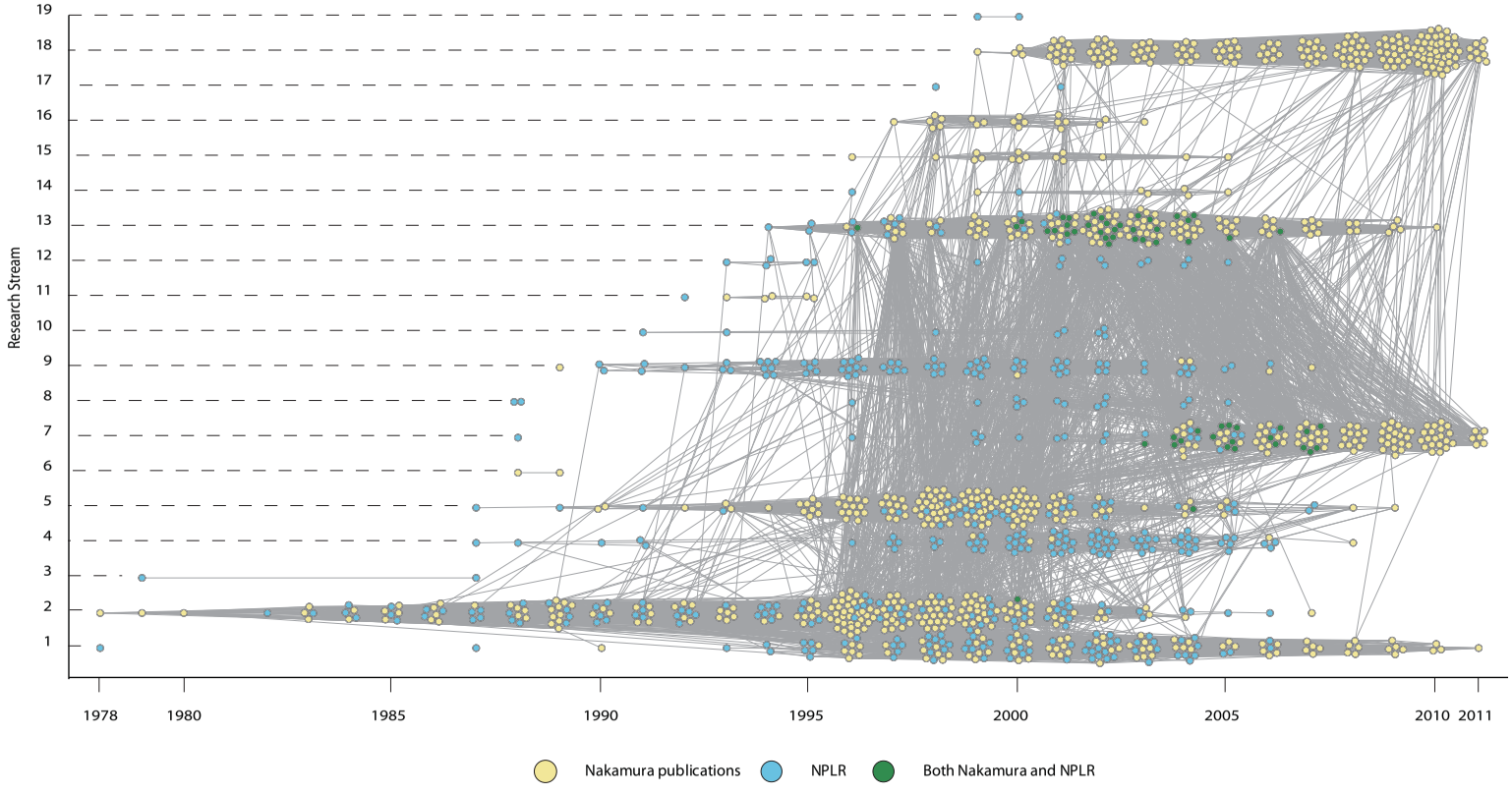
Rathenau Instituut

Figure 3 shows a similarity network of the publications and NPLRs to demonstrate the longitudinal aspects of the streams and the degree of similarity between them. Figure 3 and Table 1 reflect the same data but Figure 3 provides an overview of how the different streams are linked. The lines between the nodes represent citation relations. Figure 3 identifies Nakamura's publications, the NPLR, and papers that belong to both these categories. As table 1 shows, only two of Nakamura's research streams (7 and 13) are cited frequently by the patent applications, and two others are cited very incidentally (2 and 5). The other NPLR is located within streams that also contain (many) Nakamura publications.

Stream 18 is a large set of publications (183 in total) exclusively by Nakamura, which show little similarity to any of his previous or concurrent works. Finally, stream 7 is interesting in that the NPLRs precede the stream, with no similar publications by Nakamura. However, 5 years after the start of the stream, more and more publications by Nakamura appear and are increasingly cited by the patent applications.

The data found in Table 1 and Figure 3 are used for descriptors 1 (the reputational and applica-

Figure 3 Longitudinal research stream clustering of Nakamura and NPLR publications



Note: edges between publications signify degree of similarity based on shared combinations of title words and cited references. Nodes have been coloured to indicate source.

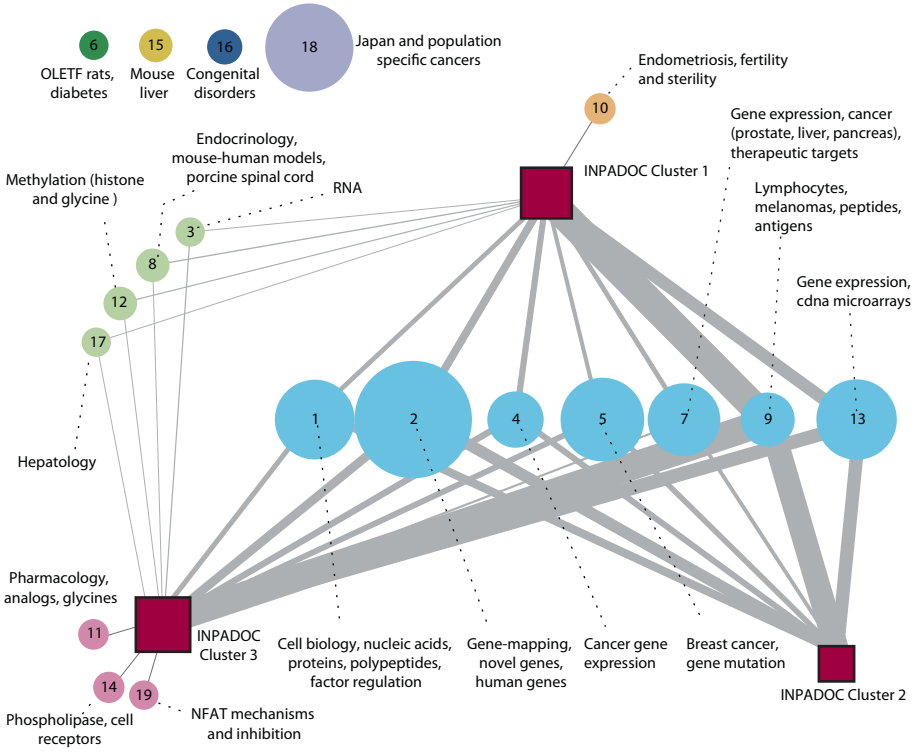
Rathenau Instituut

bility aspects) and 2 (the overall research trajectory of Nakamura overlapping with the required knowledge of the technologies). Nakamura's research is cited prominently in two research streams and co-occurs in multiple streams. His research may be considered vitally necessary in streams 7 and 13 (from Table 1), and his overall research trajectory is well embedded in the technologies (from Figure 3). There are some aspects of the technologies that do not draw extensively from Nakamura's expertise, such as those found in streams 4, 8 and 9 (from both Table 1 and Figure 3). The research streams in which Nakamura's work is cited by the patent applications are also widely cited by other publications (Table 1).

4.6.4 Patents and publications

We constructed a 2-mode network map using the research streams cited by the INPADOC family IPC clusters. This is presented in Figure 4. The publication communities in which the cited NPLRs are co-located are represented by circles, and the square nodes represent the INPADOC clusters. INPADOC clusters 1, 2 and 3 have NPLRs co-located within seven research streams; whilst INPADOC clusters 1 and 3 have four shared streams. INPADOC cluster 2 does not link to any unique publication communities.

Figure 4 Annotated 2-mode map of INPADOC family clusters and research streams



Rathenau Instituut

Note: streams not cited by patent applications are shown in the top-left corner. For INPADOC clusters, size of node indicates count of INPADOC families. For research streams, size of node indicates count of publications in stream. Edge weight is the proportional count of NPLR citations from INPADOC clusters to streams. Annotations are summarised from the most used title words and journal names/categories for each research stream.

These results address specific descriptors, namely (1) the reputational aspect (as defined by number of citations) and applicability aspect (as defined by the proportion of the researcher/inventor's work cited by the patent applications versus their total corpus) and (2) Nakamura's trajectory being co-located in the fields of research necessary for the technologies. Table 1 gives the average citations for each of Nakamura's publications. In four streams, his publications are cited over 45 times, and in seven of the other streams his publications are cited at least 25 times on average. Just over 10% of his publications are cited by the patent applications, and are found in both the IT-NPLRs and B-NPLRs. Even if Nakamura's publications are not cited by the patent applications, they are intimately co-located within the same topic and background as the NPLRs. Aspects of Nakamura's research are utilised in all three INPADOC clusters, and there are some sub-fields of his research that are utilised by only one, or two, INPADOC clusters. Other fields/sub-fields such as research streams 8, 10 and 12, are used for the technologies (as cited by INPADOC clusters 1 and 3) but are not part of, or similar to, Nakamura's research. This indicates that creating these technologies also requires knowledge that is outside of the expertise of Nakamura.

Co-inventors and partner institutes

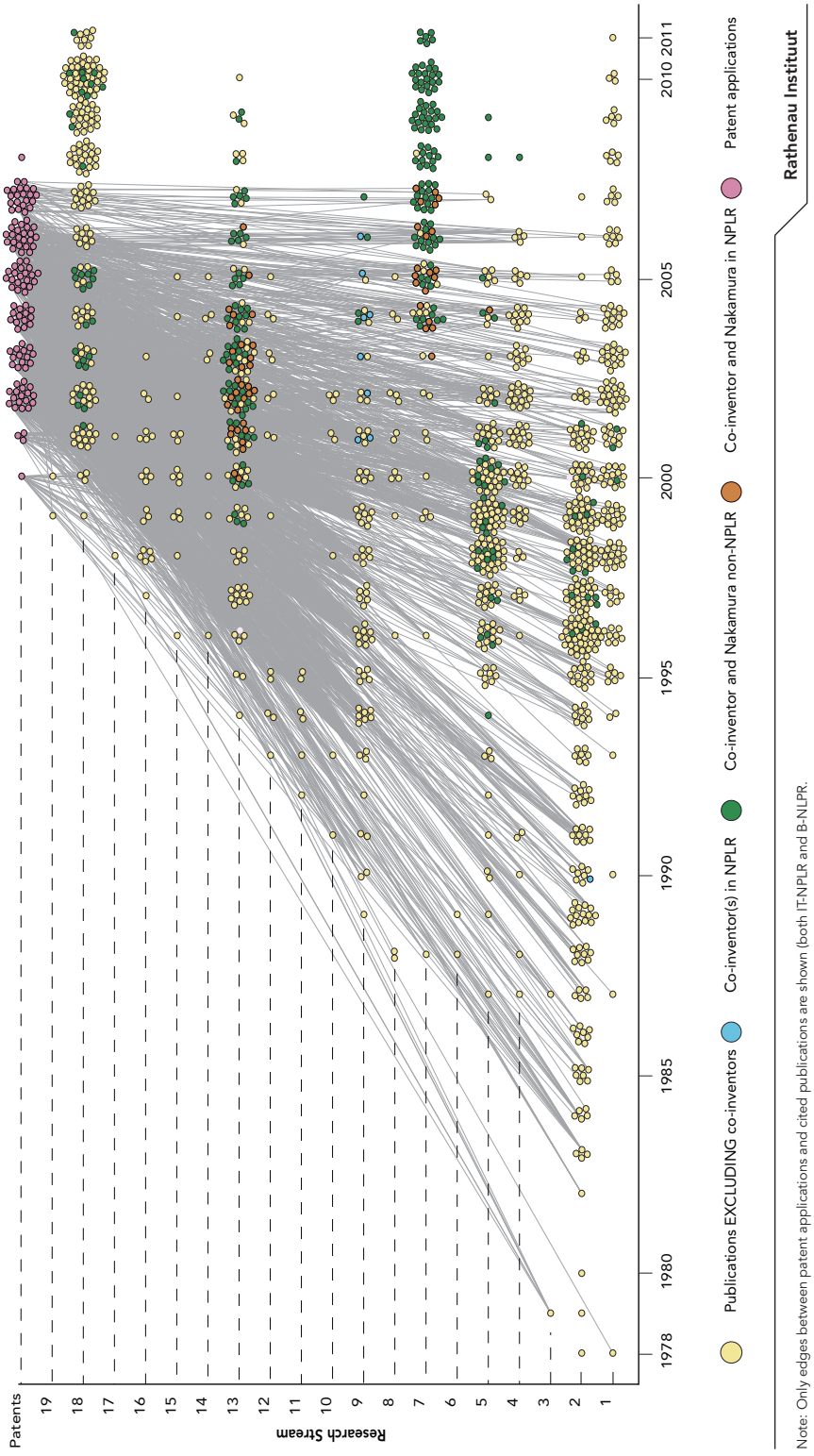
Figure 5 shows the distribution of Nakamura's co-inventors in the publication corpus. Many of Nakamura's publications are co-authored by his co-inventors. Some of the NPLR-cited publications by his co-inventors do not include Nakamura as an author. This could indicate that the knowledge utilised by the patent applications stems not only from Nakamura, but also from his co-inventors. However, the relative scarcity of NPLRs that do not include Nakamura as author but with one of his co-inventors authoring instead would suggest that the knowledge does come from within Nakamura's own research group.

Within the 77 INPADOC families that list OncoTherapy as assignee and Nakamura as inventor, Nakamura has 10 recurring co-inventors, and 4 of these co-inventors also patent without Nakamura. OncoTherapy has 6 researchers that patent without Nakamura, but the vast majority of INPADOC families primarily stem from patent applications with Nakamura listed as inventor.

OncoTherapy only collaborates on patents with two organisations: the University of Tokyo in 26 different INPADOC families, and Sentan Kagaku Gijutsu Incubation Center in 1 INPADOC family. The University of Tokyo is present in just under a third of OncoTherapy's INPADOC families, which - considering Nakamura is based at the university - does not seem particularly high.

To address the research question through descriptor (4) - the role of collaborators in the development of the technologies - we conclude that whilst there is input from Nakamura's co-inventors in the NPLR, the number of NPLRs authored without Nakamura is very low. However, we do find instances of research conducted by collaborators (stream 9 from Figure 5) necessary for all three INPADOC clusters. In this instance, whilst Nakamura does not possess the necessary expertise, his collaborators fill the necessary gap.

Figure 5 Co-inventor location in research streams



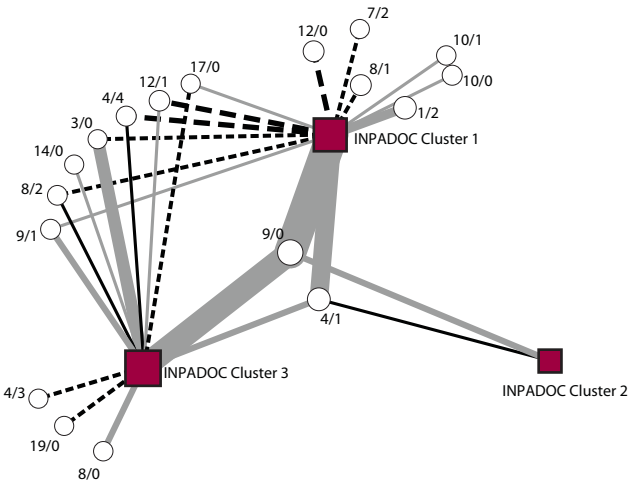
Note: Only edges between patent applications and cited publications are shown (both IT-NPLR and B-NLPR.

4.6.5 Concept clusters

From the 19 research streams, we extracted 66 concept clusters that contain NPLRs. We linked these to the citing INPADOC families and the designated INPADOC clusters. Presented in Figures 6(a)-(c) are citations to concept clusters from the INPADOC clusters. The three figures show different concept clusters being cited by the patent applications within the INPADOC clusters. Figure 6(a) shows the concept clusters that contain no publications authored by Nakamura. Figure 6(b) shows concept clusters containing publications cited by the patent applications but not authored by Nakamura, and publications authored by Nakamura but not cited by the patent applications. Finally, Figure 6(c) shows concept clusters that contain publications cited by the patent applications and are authored by Nakamura. This was done to clarify the specific research ideas and associated knowledge and skill sets that are necessary for the technologies, rather than using the broad research streams, and the presence of Nakamura's publications within these streams. The presence (or lack thereof) of Nakamura's publications within a concept cluster indicates at a specific level Nakamura's contributions to the technologies, in terms of direct citations, and through similar knowledge and skill sets. The timing of Nakamura's publications' entry into the concept clusters is also important. If Nakamura's publications were present from an early stage we can assume that the necessary knowledge for the technologies was always present within Nakamura from the beginning. Following this line of reasoning, if publications by Nakamura appear later on in the concept cluster, we assume that Nakamura realised the importance of conducting his own research in the topics that he deemed to be necessary for the technologies. We can also assume that through performing the research, he gained a greater understanding of, and thus ability to conduct, further necessary research.

Figure 6(a) shows that the patent applications located in all the INPADOC clusters rely heavily, and from an early stage, on stream 9, especially concept cluster 0 (research related to the cytotoxic effect of lymphocytes and leucocytes in human cells). Similarly, the research in stream 9 (CC9/1), cited by INPADOC clusters 1 and 3 addresses the same general topics of lymphocytes, melanomas, peptides and antigens, but focuses more specifically on the characteristics of the human leucocyte antigens (HLA) such as identification, populations and susceptibility. INPADOC cluster 1 is the only cluster to cite research from stream 7 (CC7/2, increasing rates of bile duct cancer) and from stream 1 (CC1/2, mRNA binding proteins expression and cancer proteins).

Figure 6(a) Concept clusters cited by INPADOC clusters containing only NPLR not authored by Nakamura



Rathenau Instituut

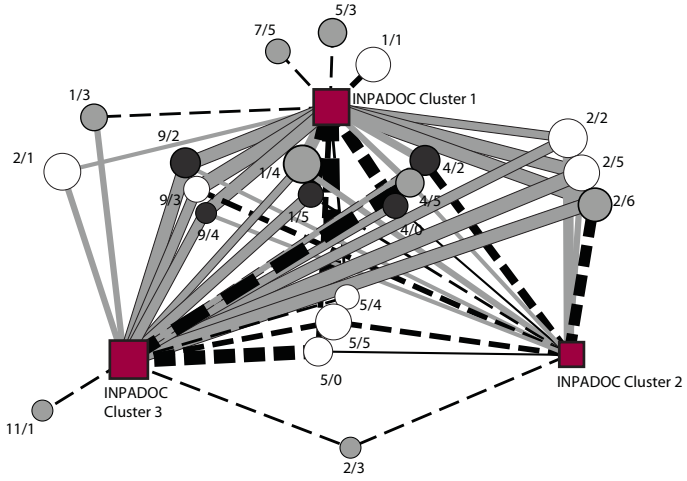
Note: For concept labels, a/b, a=parent stream ID, and b=concept cluster ID. Size of nodes=count of publications or count of INPADOC families. Thickness of edges=number of citing INPADOC families. Edge colours: age of the INPADOC cluster the concept is cited, grey=early, dashed=middle, black=late.

What is important to take away from Figure 6(a) is that the patent applications are citing research outside of Nakamura's immediate expertise. The cited publications may be providing background or introductory information (as in stream 7, CC7/2), and these publications do not link to any of Nakamura's publications through their title words and cited references.

Figure 6(b) shows concept clusters containing publications cited by the patent applications but not authored by Nakamura, and publications authored by Nakamura but not cited by the patent applications. This combination of sources indicates that there is some immediate similarity between research performed by Nakamura and the cited publications. Compared to the relative sparseness of Figure 6(a), the figure contains many more concept clusters. In many cases, the research is cited from an early stage (grey edges) but there is a fair degree of research cited later in the technologies' development phases (dashed edges).

Concept clusters from stream 9 (CC9/2, CC9/3 and CC9/4) are cited early by all three clusters, but Nakamura only starts to publish much later in these topics (indicated by shading of the concept cluster nodes). We can confirm this by examining stream 9 in Figure 3, which shows a large corpus of NPLRs, with Nakamura only appearing as author 10-15 years after the date of the first NPLR in that stream. Fig 6(b) shows that he started publishing late in most of the relevant cited concept clusters of stream 9. All three INPADOC clusters cite research in stream 1 (CC1/4 and CC1/5), but again Nakamura's publications related to those topics are only published later. Stream 1, CC1/1, is cited exclusively by INPADOC cluster 1 in the middle phase of its development (dashed edges). Notably, Nakamura - whilst having published extensively in that concept cluster - is not cited at all.

Figure 6(b) Concept clusters cited by INPADOC clusters containing NPLR not authored by Nakamura non-NPLR publications authored by Nakamura



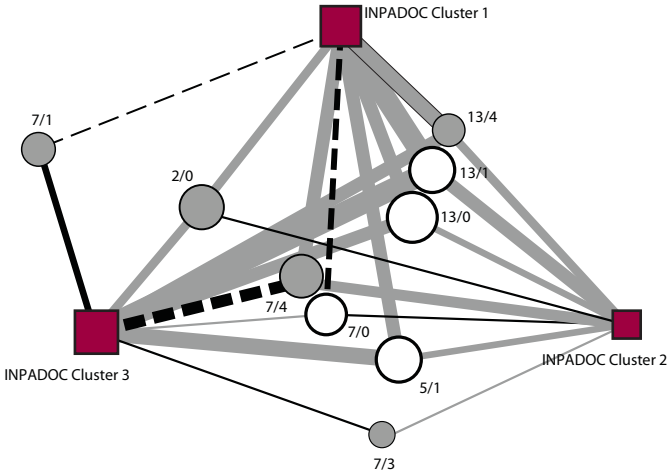
Rathenau Instituut

Note: For concept labels, a/b, a=parent stream ID, and b=concept cluster ID. Size of nodes=count of publications or count of INPADOC families. Thickness of edges=number of citing INPADOC families. Edge colours: age of the INPADOC cluster the concept is cited, grey=early, dashed=middle, black=late. CC node colours for (b) and (c): White=Nakamura publications present from start, grey=Nakamura publications present from middle time period, black=Nakamura publications present at end of time period.

Figure 6(c) shows the concept clusters in which publications by Nakamura are directly cited by the patent applications. These are considered to contain the most specifically necessary aspects of Nakamura's research for the technologies encompassed by INPADOC clusters 1-3. In most cases, the INPADOC clusters cite them from an early stage (as indicated by the grey lines) but in several other cases Nakamura did not start publishing in the topics from the beginning. This is the case with stream 13 (CC13/4), stream 7 (CC 7/1, CC7/4), and stream 2 (CC2/0). This is a strong indicator that Nakamura at some stage recognised that he had to perform his own research in these topics. As the concept clusters were cited before Nakamura began publishing in them, we attribute this observed pattern to Nakamura's ability to acquire and assimilate the required knowledge - and recognise the need to do so.

The entry period of Nakamura's publications and NPLRs to the overall publishing landscape (as found in Figures 6(b) and (c)) is important. If Nakamura is present from the beginning of the concept cluster date range (indicated by white nodes), we assume Nakamura possessed the knowledge base and skill sets as the technology was being prepared. If Nakamura entered later, we assume he did not possess the knowledge and skill sets at the time, but acquired them later.

Figure 6(c) Concept clusters cited by INPADOC clusters containing NPLR authored by Nakamura.



Rathenau Instituut

Note: For concept labels, a/b, a=parent stream ID, and b=concept cluster ID. Size of nodes=count of publications or count of INPADOC families. Thickness of edges=number of citing INPADOC families. Edge colours: age of the INPADOC cluster in which the concept is cited, grey=early, dashed=middle, black=late. CC node colours for (b) and (c): White=Nakamura publications present from start, grey=Nakamura publications present from middle time period, black=Nakamura publications present at end of time period.

The two final descriptors used to address the research question can now already be seen: (5) the degree of knowledge features utilised by the technologies that come from Nakamura in relation to other sources; and (6) the incorporated and applied skill sets acquired during the development of the technologies. We shall discuss a few typical examples: research streams 1, 7, 9 and 13.

In stream 1, Nakamura is not cited by the patent applications at all. However, he publishes extensively in the sub-field represented by the stream, and he was present at the start or early stages of all but one of the concept clusters cited by the INPADOC clusters. CC1/2 is the only part of stream 1 where Nakamura does not have a presence. This is reflected in Figure 6(a). Two (CC1/4 and CC1/5) are cited by all three clusters; one (CC1/3) is cited by clusters 1 & 3; and one (CC1/1) is cited exclusively by INPADOC cluster 1. From this information, the degree of exogenously-generated knowledge is high in stream 1, with no direct contributions by Nakamura. However, the shared knowledge base and shared minimum skill set is also high because only one of the five cited did not contain any Nakamura publications.

Stream 7 is initially hardly cited by the INPADOC clusters. Only CC7/3 and CC7/4 are cited from an early stage by INPADOC cluster 2, and CC7/0 by cluster 3 (see Figure 6(c)). Confirming this, from Figure 3, up to 2004 the first cited NPLRs were all non-Nakamura NPLRs. From 2004 onwards, almost all of the NPLRs cited in stream 7 were authored by Nakamura, and he went on to publish prolifically on that subject. Of the NPLRs that Nakamura authored, all three INPADOC clusters cite stream 7 concept clusters varyingly. Nakamura does play a direct role in the technologies of the clusters as evidenced by the proportionally large number of Nakamura-NPLRs. Nakamura's

knowledge base and skill sets, whilst being developed at a later stage have now become integral to the technologies.

The role of Nakamura in the subject areas of stream 9 is limited. He has no NPLRs in stream 9. He has a limited number of publications in all of the concept clusters cited and in two of these (CC9/2 and CC9/4) his publications appear only in later stages. Additionally, the technologies in the clusters cite the stream extensively, as seen in Figure 6(a). However, his co-inventors are cited in the NPLRs. As such, the necessary scientific aspects derived from stream 9 are outside his expertise but from within his scientific network. Stream 9 NPLRs are cited extensively by all three INPADOC clusters, suggesting that the research published in community 9 may be more than base or legacy research, i.e. research that is required to be cited by necessity in a historical development sense (an example of this is fundamental research conducted many decades prior).

As seen in Figure 3, Table 1 and Figure 6(c), stream 13 is dominated by publications authored by Nakamura. Almost a third of his publications are cited by the patent applications. From Figures 6(a)-(c), three (CC13/4, CC13/1 and CC13/0) are cited significantly by all three clusters. In one, however, (CC13/4) Nakamura is not the first to publish, with some NPLRs existing before he published in that stream.

4.7 Summary and conclusion

Professor Nakamura's past research forms the backbone of the technologies of OncoTherapy. Given that Nakamura founded the firm based on his research at the University of Tokyo, at a superficial level this is to be expected. However, the depth of his knowledge that is utilised by the firm is extensive, and that extensiveness was only found through the methodology deployed in this paper. There are aspects of the technologies that lie outside Nakamura's (and his co-inventors') expertise, and Nakamura has adopted and adapted these necessary aspects into his own research and the output of the firm. In some small sense, his co-inventors have bridged the knowledge gap between Nakamura's expertise and what is required for the technologies, but Nakamura's uptake of these research areas has filled his expertise gap.

The research question can be addressed by means of the series descriptors developed in this paper: (1) Nakamura conclusively adds to the reputational and applicability aspects of the scientific base work of the technologies; (2) His overall trajectory is closely intertwined with the fields of research necessary for the technologies, and in many cases the trajectories of both his university research and firm application have been in lock-step, with (3) fields outside his expertise initially contributing to the technologies. (4) There is a low level of input from his collaborators, with very little overlap between his co-authors and co-inventors, with (5) only some shared knowledge features shared between them. The most important aspect of Nakamura's links between his academic research and industrial applications is that (6) he incorporated skills acquired during research, in both his academic and industrial trajectories, and applied them to new research endeavours.

From the point of view of the absorptive capacity dimensions of Zahra and George (2002) and their respective sources of knowledge, be they generated exogenously or endogenously, we observed:

a. *Acquisition* where Nakamura's co-inventors were considered part of the acquisition dimension of absorptive capacity as they provided their own expertise and skill sets, which added to the knowledge infrastructure in place. Nakamura, however, in his role as Professor at the University of Tokyo, is the most active member of the firm in terms of acquisition.
b. *Assimilation* processes, where 'learning by doing' seemed to be prevalent. By conducting research in the topic areas required for the technologies, whether through a non-concerted approach (such as in exploratory research) or a cumulative directed approach (such as in a strategic, application-driven approach to research), the newly developed skills and insights impacted the development of the technologies at OncoTherapy. This suggests a high degree of assimilation by Nakamura and his co-inventors, including tentative evidence of knowledge creation feedback between Nakamura's work as an inventor and his research as an academic scientist.
c. *Transformation* and *exploitation*, where patent applications cited research outside Nakamura's expertise, but later became deeply embedded in his research trajectory. In other words, the knowledge upon which the technologies are dependent was previously externally sourced, but has been incorporated, transformed and exploited by Nakamura.

On a methodological level, our approach benefits from its ability to encompass both the macro knowledge environment and the micro knowledge capture processes. Our approach can isolate and highlight specific aspects of utilised knowledge in relation to the knowledge features already locally in place. We are able to co-locate the knowledge features of individuals who contribute to the publications and patent applications, not through the direct citations of NPLRs, but through the co-location of NPLRs in their wider knowledge environment.

A possible disadvantage of our method is the complexity of the process. Due to this complexity we chose to aggregate the technologies into clusters of INPADOC families. This limits our attention to detail within the technologies but allows a thorough examination of the contributions of an individual (in this case, Professor Nakamura). The alternative strategy would be to aggregate on the publication side and examine in detail the characteristics of the technologies being produced. This is partially done in figure 2, where we identify the relevant patent families in more detail. To do both at the same time would require more space than is possible in this publication.

Our method makes it possible to position technologies and the knowledge contributions of those involved in the development of those technologies. With the addition of funding information in the metadata extracted from WoS, it would be possible to trace the results of such funding to their exploitation. The method can also be used to examine firms that have more than one bridging scientist, and operate between multiple universities or firms. The scaling up of this method allows research groups, departments or entire research institutes or infrastructures to map their contributions in the early stages of the development of a technology right through to its exploitation or implementation. This would be useful for funding agencies and universities for reporting on their research achievements, as in many cases the end-point of fundamental and applied research may be so far removed from the origin as to be unrecognisable.

4.8 References

Ahuja, G. & Katila, R. (2001). Technological acquisitions and the innovation performance of acquiring firms: A longitudinal study. *Strategic management journal, 22*(3), 197-220.

Arundel, A. (2001). The relative effectiveness of patents and secrecy for appropriation. *Research Policy, 30*(4), 611-624.

Arundel, A. & Kabla, I. (1998). What percentage of innovations are patented? Empirical estimates for European firms. *Research Policy, 27*(2), 127-141.

Baba, Y. et al. (2009). How do collaborations with universities affect firms' innovative performance? The role of "Pasteur scientists" in the advanced materials field. *Research Policy, 38*(5), 756-764.

Blondel, V.D. et al. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*, P10008.

Bozeman, B. (2000). Technology transfer and public policy: a review of research and theory. *Research Policy, 29*(4-5), 627-655.

Braam, R.R. et al. (1991). Mapping of science by combined co-citation and word analysis. I. Structural aspects *Journal of the American Society for Information Science and Technology 42*(4), 233-251.

Breschi, S. et al. (2003). Knowledge-relatedness in firm technological diversification. *Research Policy, 32*(1), 69-87.

Callaert, J. et al. (2006). Traces of prior art: An analysis of non-patent references found in patent documents. *Scientometrics, 69*(1), 3-20.

Callon, M. et al. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry. *Scientometrics, 22*(1), 155-205.

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

Cohen, W.M. et al. (2002). Links and impacts: the influence of public research on industrial R&D. *Management Science*, 1-23.

Courtial, J.P. et al. (1993). The use of patent titles for identifying the topics of invention and forecasting trends. *Scientometrics 26*(2), 231-242.

Criscuolo, P. & Verspagen, B. (2008). Does it matter where patent citations come from? Inventor vs. examiner citations in European patents. *Research Policy, 37*(10), 1892-1908.

Engelsman, E.C. & van Raan, A.F.J. (1994). A patent-based cartography of technology. *Research Policy, 23*(1), 1-26.

Fischer, M.M. (2001). Innovation, knowledge creation and systems of innovation. *The Annals of Regional Science, 35*(2), 199-216.

Fleming, L. & Sorenson, O. (2001). Technology as a complex adaptive system: evidence from patent data. *Research Policy, 30*(7), 1019-1039.

Furukawa, R. & Goto, A. (2006). Core scientists and innovation in Japanese electronics companies. *Scientometrics, 68*(2), 227-240.

Garfield, E. & Welljams-Dorof, A. (1992). Citation data: their use as quantitative indicators for science and technology evaluation and policy-making. *Science and Public Policy, 19*, 321-321.

Gorman, M.E. (2002). Types of knowledge and their roles in technology transfer. *The Journal of Technology Transfer, 27*(3), 219-231.

Griliches, Z. (1998). *Patent statistics as economic indicators: a survey*: University of Chicago Press.

Gurney, T. et al. (2012). *Knowledge Capture Mechanisms in Bioventure Corporations*. Paper presented at the 17th International Conference on Science and Technology Indicators (STI), Montreal.

Hicks, D. (1995). Published papers, tacit competencies and corporate management of the public/private character of knowledge. *Industrial and Corporate Change, 4*(2), 401.

Horlings, E. & Gurney, T. (2012). Search strategies along the academic lifecycle. *Scientometrics*, 1-24.

Hullmann, A. & Meyer, M. (2003). Publications and patents in nanotechnology. *Scientometrics, 58*(3), 507-527.

Jaffe, A.B. (1986). Technological Opportunity and Spillovers of R & D: Evidence from Firms' Patents, Profits, and Market Value. *American Economic Review 76*(5), 984-1001.

Karki, M. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information, 19*(4), 269-272.

Lanciano-Morandat, C. et al. (2009). Le capital social des entrepreneurs comme indice de l'émergence de clusters ? *Revue d'économie industrielle*(4), 177-205.

Leydesdorff, L. (2008). Patent classifications as indicators of intellectual organization. *Journal of the American Society for Information Science and Technology, 59*(10), 1582-1597.

Martinez, C. (2010). *Insight into different types of patent families*: OECD.

McMillan, G. et al. (2003). The impact of publishing and patenting activities on new product development and firm performance: the case of the US pharmaceutical industry. *International Journal of Innovation Management, 7*(2).

Meyer, M. (2000). Does science push technology? Patents citing scientific literature. *Research Policy, 29*(3), 409-434.

Meyer, M. (2002). Tracing knowledge flows in innovation systems. *Scientometrics, 54*(2), 193-212.

Meyer, M. (2007). What do we know about innovation in nanotechnology? Some propositions about an emerging field between hype and path-dependency. *Scientometrics 70*(3), 779-810).

Meyer, M.S. (2001). Patent citation analysis in a novel field of technology: An exploration of nano-science and nano-technology. *Scientometrics, 51*(1), 163-183.

Narin, F. (1976). Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity. *Cherry Hill, NJ: Computer Horizons, 338*, 27.

Narin, F. (1994). Patent bibliometrics. *Scientometrics, 30*(1), 147-155.

Narin, F. (1995). Patents as indicators for the evaluation of industrial research output. *Scientometrics, 34*(3), 489-496.

Nelson, A.J. (2009). Measuring knowledge spillovers: What patents, licenses and publications reveal about innovation diffusion. *Research Policy, 38*(6), 994-1005.

Nelson, R. (2004). The market economy, and the scientific commons. *Research Policy, 33*(3), 455-471.

Owen-Smith, J. & Powell, W.W. (2003). The expanding role of university patenting in the life sciences: assessing the importance of experience and connectivity. *Research Policy, 32*(9), 1695-1711.

Pavitt, K. (1988). Uses and abuses of patent statistics. *Handbook of quantitative studies of science and technology*, 509-535.

Ponomariov, B. & Boardman, C. (2012). *Organizational behavior and human resources management for public to private knowledge transfer: an analytic review of the literature*: OECD Publishing.

Rosenberg, N. (1990). Why do firms do basic research (with their own money)? *Research Policy, 19*(2), 165-174.

Schartinger, D. et al. (2002). Knowledge interactions between universities and industry in Austria: sectoral patterns and determinants. *Research Policy, 31*(3), 303-328.

Schmoch, U. (1993). Tracing the knowledge transfer from science to technology as reflected in patent indicators. *Scientometrics, 26*(1), 193-211.

Schmookler, J. (1966). *Invention and economic growth*: Harvard University Press Cambridge, MA.

Somers, A. et al. (2009). *Science Assessment Integrated Network Toolkit (SAINT): A scientometric toolbox for analyzing knowledge dynamics*. The Hague: Rathenau Instituut.

Tijssen, R.J.W. (1998). Quantitative assessment of large heterogeneous R&D networks: the case of process engineering in the Netherlands. *Research Policy, 26*(7-8), 791-809.

Tijssen, R.J.W. (2001). Global and domestic utilization of industrial relevant science: patent citation analysis of science-technology interactions and knowledge flows. *Research Policy, 30*(1), 35-54.

Tijssen, R.J.W. (2002). Science dependence of technologies: evidence from inventions and their inventors. *Research Policy, 31*(4), 509-526.

Tijssen, R.J.W., & van Raan, A.F.J. (1994). Mapping changes in science and technology. *Evaluation Review, 18*(1), 98-115.

van den Besselaar, P. & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: a method and an exploratory case study. *Scientometrics, 68*(3).

White, H.D. & McCain, K.W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science (1986-1998), 49*(4),327-355.

Zahra, S.A. & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. *Academy of Management Review, 27*(2) 185-203.

Zucker, L.G. & Darby, M.R. (1996). Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry. *Proceedings of the National Academy of Sciences of the United States of America, 93*(23), 12709.

Zucker, L.G. et al. (1999). *Intellectual capital and the birth of US biotechnology enterprises*: National Bureau of Economic Research.

5 Social and Scientific Networks of Founders of Start-Ups at Leiden Bioscience Park[1]

Abstract

An idea generated in academia and exploited in industry travels along a complex pathway. Science Parks aim to help start-ups that exploit the skills and knowledge of fellow tenants in a Science Park and also further develop those that they acquired in previous research at the university. In this study, we analyse the technology development pathways of single-site start-ups located at Leiden Bioscience Park, and the access to resources in the physical and social environments that it requires. We conduct interviews with the founders of the start-ups, examining sources of social capital in developmental phases of the start-ups. We examine the social and physical proximity of the firm and firm founders to stakeholders inside and outside the Science Park, including the contextual origins and application spheres of the sources of social capital available to start-ups prior to, and after, their choice to locate within a Science Park. To examine the technology development pathways, we link the patent applications of the firms to the firm founders' publications through the non-patent literature references that are most similar to their publication corpora. We find that the sample set of firms integrate new streams of academic research, primarily from their alma mater, into their technological output, in addition to continuing, and expanding upon, their own research streams. The relationship with the local university (if different from its original affiliation) increase too. Apart from that, the social capital utilised by the firms comes from outside the Science Park with minimal involvement from the Science Park administration or other firms located within the Science Park.

5.1 Introduction

From inception to exploitation, a quantum of knowledge follows a convoluted route. In the most simple of models, inputs lead to outputs through a black box of context, processes, skills and previous knowledge (see for example Autio et al. 2004), supported by the infrastructures required to host the processes and skills. The infrastructures derive from numerous policy, education, and innovation environments. The Science Park is one of these infrastructures.

Science Parks have entered the literature in waves with each crest bringing new ideas and theories as to their utility to science, innovation, and society. Studies on Science Parks most frequently use a variety of methods and approaches - including questionnaires, interviews, financial data, patent data, and more - to evaluate the utility of a Science Park (Dettwiler et al., 2006), to compare Science Parks (Fukugawa, 2006), or to compare firms on and off Science Parks (Squicciarini, 2008). In most of these studies, the knowledge capture black box remains closed.

1 This chapter will be published as Gurney, T. et al. (2013) *Access and utilisation of social capital in knowledge transfer* In proceedings of Science and Technology Indicators (STI), Berlin, and as Gurney, T. et al. (2013) *From inception to exploitation: research trails in biotechnology start-ups* (2013) In proceedings of Science and Technology Indicators (STI), Berlin. This chapter has been submitted for publication by Technovation.

The transformation from exploratory research to exploited artefacts, processes and services is the end-goal of the Science Park firm, and by extension, of the host Science Park. Without this transformation, all the economic, technological, scientific and social benefits a Science Park purportedly offers are moot. Finding out what actually happens around firms on Science Parks requires us to trace this process in detail, keeping in mind that the process of transformation starts before the firm has been incorporated and before it has located on the Science Park.

In this paper we examine in detail the knowledge capture and transformation mechanisms around firms on Science Parks from the very start to the very end. We trace the social and cognitive routes of knowledge generated in academia and exploited in industry by the firm founder a Science Park environment. The firms we examine are start-up firms based at Leiden Bioscience Park. The paper is structured as follows: the next section develops our conceptual framework in relation to existing literature. Following this, the specific aims and research questions are discussed in detail. We then discuss the methodology and present our results. Our conclusions and discussions follow, including implications for further analyses and policy.

5.2 Conceptual framework

Science Parks have been wielded as a policy tool for many years, and numerous policy initiatives such as the EU Framework Programmes and the Bayh-Dole Act (which signalled a change in the intellectual property regime in favour of universities) have incentivised the formation of Science Parks across the globe (Siegel, 2003). On a regional and national innovation level, the fear of being "left behind" in technological progress has in part led to policy being enacted and Science Parks being formed (Shearmur & Doloreux, 2000).

5.2.1 Science Parks

Identifying and studying the host of development and governing processes is difficult at best. From the highest level of aggregation, the Science Park, working down in scale to the academic researcher or soon-to-be firm founder, we can identify common threads linking the levels.

1. No common definition: Science Parks, and the utility of Science Parks, have been extensively studied, yet common definitions are hard to come by. General descriptions of a Science Park amount to a property-based, technology-orientated agglomeration of firms of varying specialisations and sizes, with close links and opportunities - either cognitive, geographical, structural or commercial - between firms and to a higher education or research institution (Das, T.K. & Teng, 1997; Löfsten & Lindelöf, 2005; Quintas et al., 1992; Siegel et al., 2003). In Asia, the preferred nomenclature is 'Technology Park' whereas in North America 'Research Park' is preferred. Europe is the dominant user of the 'Science Park' term (Link & Scott, 2007).

2. Unique origins: Each Science Park comes from unique origins. Kobe Science Park was developed as a regional rejuvenation effort after the 1995 earthquake. Silicon Valley was the result of commercial agglomeration effects. Hsinchu was established by the Taiwanese government to lure back all those who had previously opted for Silicon Valley. Each park has its own unique origins and context. Some have been developed for the infrastructure, whereas others have been developed to improve R&D innovation and production, or to provide intellectual development (Koh et al., 2005).

3. Host of motivations for Science Park formation: In terms of general motivations, the most cited reasons for Science Park formation are to foster the creation and growth of R&D-intensive firms; to provide an environment for large firms to develop relationships with small firms; to promote formal and informal links between firms, universities and other small labs (Das, T.K. & Teng, 1997; Löfsten & Lindelöf, 2005; Siegel et al., 2003); to provide a contact space between "fast applied science" and "slow basic science" (Quintas et al., 1992); to promote foreign investment and accelerate transition from a labour-based economy to a knowledge-based economy (Koh et al., 2005); and to provide technological development and renewal on a regional or national basis (Castells & Hall, 1994; Felsenstein, 1994; Phillimore, 1999).

4. A Science Park must seek tenants, regardless of the ulterior motives of the firm founder. The space within the park must be filled to provide the economic and social rate of return to investments expected from a Science Park (or any large-scale research infrastructure for that matter). As noted by Phan et al. (2005), all Science Parks essentially compete with each other to attract new firms to their location, as new successful firms form the life-blood of a Science Park. Firms choosing to locate on a Science Park come either in the form of a HEI spin-off/start-up or as the subsidiary of an outside firm (without links to the HEI) that believes it is necessary to be located on a SP for various reasons.

5. Different tenants seek different benefits. In the case of a university spin-off/start-up, support structures must be in place to ensure a profitable transfer of university-originated technology, be it through licensing or manufacturing (Clarysse et al., 2005). For university and non-university originated firms, various location theories attempt to explain the behaviour of firms and the motivations they give. Neo-classical theory focuses on transport, labour costs, distances and agglomeration economies whereas behavioural theories address mediators, gatekeepers, information channels and reputational advantages. Structuralist theories deal with the innovative milieu as well as agglomeration effects due to the geographical characteristics of the locale (Westhead & Batstone 1998, Barney 2001).

Access to networks, touted by most proponents of Science Parks, can be seen as paramount for new firms locating to a Science Park. The network benefits of a Science Park can be described in terms of access to scientific and technical expertise, providing an environment for large firms to develop relationships with smaller firms, and to promote formal and informal links between firms and universities and other smaller labs (Das, T.K. & Teng, 1997; Löfsten & Lindelöf, 2005; Siegel et al., 2003). The benefits can also be financial, promoting access to investment (Koh et al., 2005); commercial, providing access to potential clients within the park; and organisational, deriving from the Science Park administration itself and from the non-scientific and technical expertise and services that other firms may bring.

Westhead and Batstone (1998) surveyed matched pairs of in-SP and off-SP firms on their motivations for choosing their location. Their results suggest that the traditional selling points championed by the developers of Science Parks do not correlate well with the actual motivations of the firms who locate there. Science Park firms assign higher importance to the prestige of a Science Park, to car parking facilities and to the fact that the key founder lives locally than to the proximity of or links to a HEI. Moreover, they find that the typical firm entering a Science Park is

more likely to be an older, longer-established firm than a recently created spin-off. Other studies do find a more positive correlation between the close presence of a HEI - with the human capital it represents - and the motivation to locate to a Science Park (e.g. Dettwiler et al. 2006).

5.2.2 Resources and networks

The literature relating to the resources and networks available to firm founders is of particular interest. These derive from areas investigating human and social capital (Adler & Kwon, 2002; Audretsch et al., 2005; Cainelli et al., 2007; Lanciano-Morandat et al., 2009; Landry et al., 2002), strategic alliances (Das, T.K. & Teng, 2000; Deeds & Hill, 1996; Parise & Henderson, 2001) and entrepreneurial development (Ho & Wilson, 2007; Murray, 2004; Oliver, 2004).

Social capital

The development of a firm founder's social capital can best be described as a supplementary, enabling resource - in addition to the stock knowledge, financial capital and skills of an entrepreneur (Dubine & Aldrich, 1991; Greve & Salaff, 2001; Lin, 1999). It is argued that there are two forms of social capital: bonding social capital and bridging social capital (De Carolis & Saparito, 2006). Bonding social capital refers to the ties within a network, and their effect on the norms and behaviour of the actors within that network (Adler & Kwon, 2002). Bridging social capital refers more to private benefits that an individual may gain through access to a network (Leana & Van Buren, 1999). It is the second definition of social capital that is of most interest to us in this paper.

Entrepreneurial activity is often marked by the ability of a firm founder to mobilise such bridging social capital through their familial and social ties, as well as the professional relationships they develop upon entry to a field. Initially, the professional network of an academic firm founder is based upon his or her research and environment. For founders of academic spin-offs this equates to their contemporaries and host university. As a spin-off develops, its sources of social capital begin to evolve. The priorities of the firm founder in securing and mobilising social capital change depending on the development stage of the spin-off. Entering into new social or professional networks grants access to, and interactions with, a wider variety and number of potential stakeholders and support entities, which may increase the enabling aspect of social capital (Lanciano-Morandat et al., 2009)

For firms choosing to locate to a Science Park, social capital takes on a physical and market proximity aspect (Sorenson, 2003). Science Parks draw in firms from similar markets and potential requirements for a fledgling firm can be found in the experience and capacities of these more established firms. The nearby presence of a university also opens up the possibility of entraining available academic social capital (as well as the vast human capital that the university represents). If the university is the alma mater of the firm founder, the accrued social capital can be easy to access as the networks the founder was once part of are most likely to still be in place.

A general model relating to social capital is that of Elfring & Hulsink (2003). According to Elfring & Hulsink, there are three dominant processes in developing social capital. The first, the *discovery of opportunities* is affected by prior knowledge and information about the opportunity. *Securing resources* is the second process, in which the start-up accesses, mobilises and deploys resources. The last process, *obtaining legitimacy*, involves enhancing their visibility through affilia-

tions, alliances and networks, but also through the development of the scientific 'face' that a company exposes to the market. This could include prominent links to universities, or noted professors on the advisory board.

Knowledge capture

Access to resources, assets and capabilities (Rothaermel & Deeds, 2004) includes not only the contextual and network benefits of a Science Park but also the skills and knowledge of the personnel of the firm - the human capital. The human capital of the firm begins with, and is largely shaped by, the firm founder (Bozeman et al., 2001; Schartinger et al., 2002). The firm founder can be seen as the progenitor of the firm and its technologies, for the research conducted prior to the firm's formation leads to the exploitation of that knowledge. The founder's research decisions in the years prior to forming the firm have been influenced in two stages, first when the founder was an academic researcher, and second when the founder became an industrial researcher.

The exploration-exploitation model (March, 1991) is applicable to all business activity and has been applied widely, including in biotechnology-orientated studies. The model generally describes the need to achieve a balance between a firm's research activities (exploration) with a firm's product development and sales (exploitation). Too much emphasis on exploration can potentially increase uncertainty and risk, whilst too much emphasis on exploitation leads to a reduction in knowledge variation. Important to the practices of the exploration-exploitation model are the precursors. The precursor to exploration is simple desire or curiosity, and the precursor to exploitation is the presence of applicable resources, be they financial, economic or human (Rothaermel & Deeds, 2004). Gupta et al. (2006) address what exactly is meant by exploration-exploitation: are they orthogonal or continuous processes? Should the practitioner be good at both, be excellent at one, or is there a punctuated equilibrium between the processes?

The interactions of the firm founder with other entities within the Science Park are governed by this model. The base knowledge stock, including the results of previous exploration activities, needs to be exploited. The decision to exploit this stock is complex, and becoming an "entrepreneurial scientist" (Oliver, 2004) - managing the academic research as well as the development of the firm - is a difficult process. New concerns such as venture capital and intellectual property become more entwined with their research activities (Ho & Wilson, 2007). Whilst the scientist may not be new to the world of intellectual property or securing funding (albeit in an academic rather than a venture capital setting), from the scientist's point of view the processes operate within a different incentive structure.

Prior to the inception of a technology, the environment of the future firm founder is academia. Academia could ostensibly be called the 'proving grounds' of a start-up, where the knowledge required for the future technology and the firm (which is based on that technology) is explored and vetted. In the initial phases of research, the scientist is influenced by the reward system of science, as proposed by Merton (Merton, 1957, 1969). With a mind to the incentives structure, the scientist employs a number of strategic decisions (Horlings & Gurney, 2012), guided by the perception of the trade-off between field crowdedness, problem difficulty and potential reputational gains (Hagstrom, 1974; Zuckerman, 1992; Zuckerman & Cole, 1994).

There are multiple threads of research collectively building on the reward system of science. These include the distinction between rank and file scientists and star scientists (most notably covered by Zucker and Darby (1996) and Zuckerman (1992)), the identification via output and productivity in terms of age (Costas et al., 2010), and at the academic institute level, the life cycle effect on productivity (Levin & Stephan, 1991) and the academic life cycle of university researchers (Horlings & Gurney, 2012).

Exploration, or research for research's sake, is generally uncommon in the realm of start-ups or spin-offs. The exploitation of such research is, however, a more common practice and with public and financial pressures on science and firms for meaningful results, very much a necessity. The eventual desired output of these complex processes may not even materialise, and the processes affecting each step can contribute to this uncertainty.

5.3 Aim

We have previously developed methodological tools that describe the scientific and social aspects of knowledge capture mechanisms. The first describes the search phases of an academic researcher, included in Horlings & Gurney (Horlings & Gurney, 2012). In this model, researchers were found over the course of their career to work in a number of research trails and work simultaneously in different research trails. Also found were scientists' roles in problem selection change with age and qualification and that entry and exit from research trails was linked to potential reputational gains. An extension of that model (Gurney et al., 2012) is that the exploitation (patenting) activities can be linked to previous research. This was done through the references cited by the patent applications, and how they link in context and scientific background to the work of the publishing and patenting output of the firm founder. With this model and extension, the development over time of the scientific aspect of an idea generated in academia and exploited in industry was made visible, with regards to the selection processes and decisions of the researcher.

For the social aspect, specifically related to firms located in Science Parks, we previously researched (Lanciano-Morandat et al., 2009) the development of the social networks of biotechnology-oriented firm founders. In this study - in line with Elfring & Hulsink (2003) processes for developing social capital - we examined, through in-depth interviews with firm founders, the development of each firm through the lens of which entities (be they individuals, other firms, organisations, financial institutions and more) the founders made and maintained contact with over the current lifespan of the firm. This research compared the firm founders' resource priorities (in terms of finding and securing resources) during the development of their firms, as well as the legitimacy benefits gained by collaborations with other firms. The analysis of the interviews involved categorising the responses from firm founders into five distinct contextual spheres, various actor-types, and three time phases in a firm's development, which shall be expanded upon in the methodology section.

Following the same methodology and with the theoretical models at hand, we aim to answer the following questions: *What are the cognitive routes and developments of an idea generated in academia and exploited in industry? What relations support knowledge capture and transformation, particularly in regards to the social capital of firm founders? What is the role of Science Parks in facilitating these?* With these questions and methodological tools, we focus on and investigate:

a. The links between the firm founders' knowledge stocks and their technological output including the scientific and technological links to higher education institutions and public research facilities.
b. The continuity of research conducted by the firm founder - prior to, and at certain periods after, incorporation;
c. The composition of academic and industrial collaborations of the firm and firm founder including the regionalism/internationalism of his/her collaborators;
d. The social interactions in differing contextual settings between firm founder and stakeholders within, and outside, a Science Park;
e. The social interactions with, or mediated by, the Science Park administration.

5.4 Data and Method

5.4.1 Science Park and firm selection

Leiden BioScience Park (LBP) is the subject of our analysis. LBP is a biomedical science cluster in the Netherlands. Since its foundation in 1984, the park has grown significantly and currently has close to 100 firms located on the premises. The grounds of the park include property of both Leiden University and the Leiden University Medical Centre. The Hogeschool Leiden (Leiden University of Applied Sciences) is also located on the same premises along with other knowledge institutions such as TNO and Top Institute Pharma.[2] There is a park administration that actively searches for new firms to locate to the premises and is partnered with Biopartner, which manages the facilities at many of the premises, as well as providing advice and funding opportunities for firms located in the park.

Firms were selected on 3 primary criteria: firm formation was within the last 10 years; the firm was founded by a university or knowledge institute researcher; and, lastly, the firm is from the life sciences and health sector. Following these criteria, we were able to interview and collect full patent and publication data for 9 firms. These criteria were deployed so as to ensure certain commonalities i.e. economic climate, scientific field, approximate qualifications of the firm founder and formation origins (specifically academic spin-offs rather than corporate spin-offs).

5.4.2 Interviews

Interviews were conducted with the founders of the 9 firms based at Leiden BioScience Park. The overall aim of these interviews was to elucidate the communication linkages between the firm founders and various actor types from varying sectors. The interviews were semi-structured with a pre-determined list of topics to be discussed. If any of the topics were not discussed in the interview, they were asked as direct questions at the end of the interview. The topics revolved around the nature of interactions between the firm founder and stakeholders involved during the development of the firm. The topics and interview coding typology (from Lanciano-Morandat et al. (2009)) concerned:

2 For more information about Leiden Bioscience Park, www.leidenbiosciencepark.nl/fact_sheets

1. The origins of the stakeholder:
   - Academia (e.g. universities, scientific advisors or students)
   - Organisational and training groups (e.g. patient groups, consortia or professional networks),
   - Finance (e.g. banks or venture capital firms),
   - Commerce (e.g. customers, marketing firms or suppliers)
   - Industrial partners (e.g. manufacturers, other biotechnology firms or pharmaceutical partners)
   - Policy and regulation (e.g. lawyers, trial administrators/enablers or safety officials)
   - Science Park (e.g. Science Park administrators or facilities management).

2. The specific context of the interaction which included:
   - Scientific - the scientific knowledge involved in developing the firms' products or processes;
   - Commercial - the commercial or sales aspect of the products or processes offered by the firm;
   - Financial - the funding of research through venture capital or grants etc;
   - Technical - the technical (including equipment) knowledge required for the functioning of research activities;
   - Organisational - the regulatory and/or administrative requirements for product/process development or for firm operation.

3. The proximity of the interacting individual/institution of which the entity could be:
   - Personally related - in which the entity is/was a family member, friend or close acquaintance known before the firm was founded;
   - Not personally related but within the physical confines of the Science Park;
   - Not personally related but from outside the physical confines of the Science Park.

4. The interactions are classed according to phases in the formation of the firm, specifically:
   - The pre-entrepreneurial phase (this includes the time before the firm was officially incorporated, to shortly after incorporation);
   - Entrepreneurial phase (wherein the technology of the firm was believed to have passed its viability phase);
   - Managerial phase (where the duties of the founder as CEO have migrated to that of CSO).

Due to the content and depth of the issues discussed with the interviewees and in accordance with confidentiality agreements with the interviewees/firms, the results presented have been generalised with all identifying data removed.

The interviews were transcribed and analysed through careful reading of the transcripts (along with notes taken during the interview). We categorised the entities with which the firm founder had contact according to their origin and the environment in which they operate. We noted the type and extent of any interactions between the founder and the entity, along with the period in the lifespan of the firm, to create matrices presenting the four typologies, presented above.

5.4.3 Patents and publications

For patent data we use the PatSTAT database prepared and developed by the EPO. We extracted all patent applications with the firm or firm founder listed as an applicant, or the firm founder

listed as inventor. Variations of the names used as search input were included and the results were manually cleaned. If any discrepancies remained, we put these directly to the firm founder.

For publication data, we used Thomson Reuters' Web of Science (WoS) as our primary source, supplemented by CV data from the scientists involved. All publications by the firm founder were downloaded from WoS (all entries up to June 2012).

These base data were parsed using SAINT (2009) and managed in a relational database. Further data were collected from the patents, specifically:

1. In-text non-patent literature references (IT-NPLRs) - citations to publications visible in the body of the patent.
2. Bibliographic NPLRs (B-NPLRs).

Both NPLR sets were parsed and, as far as possible, their WoS publication equivalents retrieved. A manual check was performed to see whether the retrieved documents matched the original NPLR. If any discrepancies in metadata were found, we used the WoS version or the most common usage of the specific data point, but if any discrepancies remained, we did not use the records for any further analysis. Examples of modifications to metadata in this process included publications cited by patent applications using one year, but the matching publication in WoS using another year. In this instance, the data within WoS was taken to be the correct data. The verified documents were then parsed and processed separately for a firm-specific analysis and collectively for a group analysis. We coded the addresses found within the publication and patent application data by country of origin and type of entity, including individuals, university/public research organisations or industrial entities.

Patent and publication visualisation and analysis

The patent applications and scientific publications were grouped together using methods by Horlings & Gurney (2012) and Gurney et al. (2012). Publications were clustered by their shared combinations of title words and cited references (van den Besselaar & Heimeriks, 2006). The degree of similarity was calculated using the Jaccard similarity index. Clusters of publications were automatically assigned by a community detection algorithm (Blondel et al., 2008) within SAINT. This algorithm groups publications based on their degree centrality and the relative weights of edges between nodes.[3] The NPLRs of the patent applications were included in the clustering of the publications and served the purpose of linking the NPLRs to the founder publications through their content and scientific background.

Using NPLRs we can link publications to patent applications, thus establishing the scientific relevance (as determined by the applicant and examiners) of the patent applications to the corpus of publications of the founder. Even if the patent applications do not directly cite the work

3 For a more detailed explanation of clustering algorithms in general, see Palla, G., Derényi, I., Farkas, I. & Vicsek, T. (2005). *Uncovering the overlapping community structure of complex networks in nature and society.* Nature, 435(7043), 814-818. For a comparative analysis of Blondel et al.'s algorithm versus others see Lancichinetti, A. & Fortunato, S. (2009). *Community detection algorithms: a comparative analysis.* Physical Review E, 80(5), 056117.

of the founder, the NPLR that are cited cluster within their corpus, inferring a link to the founder's areas of expertise. These result in a tangible, visible, shared knowledge base between the patent and the publication. This allows us to observe and elucidate an indication of the degree of knowledge transfer from the research practices and results of the founder to their technological output.

5.5 Results

5.5.1 Social interactions between firm founder and other entities over time

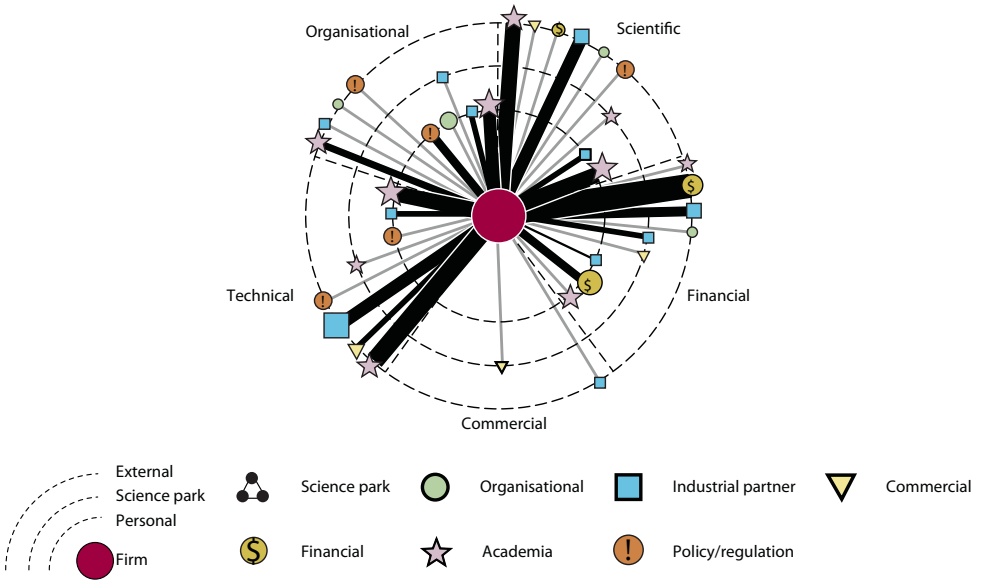
The social interactions recorded allow us to determine from where the founder gathers which resources. These resources include scientific, technical, organisational and managerial expertise, funds (both private and public) and network partners. Each domain in which the founder interacts provides specific resources determined by their context and proximity.

Pre-entrepreneurial phase

The collective interactions of the nine firms interviewed at Leiden BioScience Park are presented below. Figure 1 corresponds to the firms' pre-entrepreneurial phase (before incorporation). In terms of the firms' average interaction count (left), the greatest part of the interactions are external, followed by personal interactions and those within the Science Park. For the types of interactions reported by firms, the only interaction to be mentioned by all 9 firms was to external financial actors. The interaction reported by the most firms was with industrial partners in the technical sphere external to the Science Park.

The scientific sphere is the main origin of scientific and technical knowledge for the firm founder, primarily from external actors. There are some personal interactions to academic actor-types, and these consist of mostly scientific advice to the firm founders during the pre-entrepreneurial phase.

Figure 1 Leiden collective interactions during pre-entrepreneurial phase



Rathenau Instituut

Note: Size of node indicates average number of interactions per firm. Edge thickness signifies count of firms reporting interactions. Grey edges signify only one firm reporting interaction.

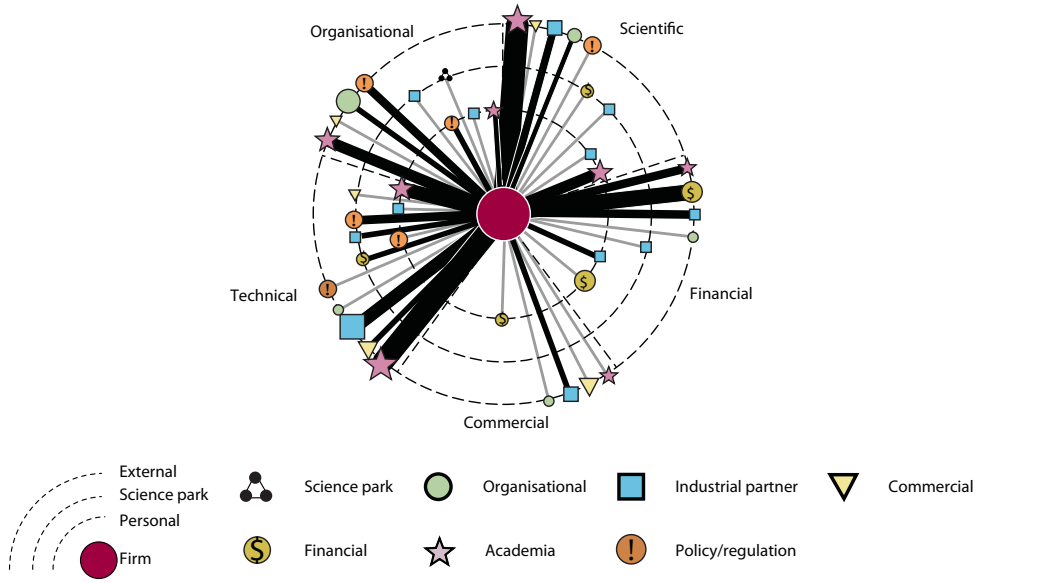
In terms of funding sources, there are only incidental interactions with financial actor types in the scientific sphere, with those being small financial interactions (such as rent payments) with universities. The majority of interactions over funding resources come from the financial sphere, and are from internal and external interactions.

In terms of proximity, personal interactions are primarily knowledge-based or financial, along with some interactions with industrial partners. External interactions are primarily related to funding sources, industrial partners and knowledge sources.

Entrepreneurial phase

Figure 2 shows the interactions reported by the firms during their entrepreneurial phase. Compared to the pre-entrepreneurial phase, the distribution of actor-types is similar, but the addition of a new actor-type, specifically the Science Park as an entity, was reported. There was an increase in the number of commercial interactions, as firms were securing their first customers. An increase was also seen in the number of interactions with academic actor types within the technical sphere. This was reported by the firms to be the result of collaborations with industrial partners.

Figure 2 Leiden collective interactions during entrepreneurial phase



Rathenau Instituut

Note: Size of node indicates average number of interactions per firm. Edge thickness signifies count of firms reporting interactions. Grey edges signify only one firm reporting interaction.

In terms of knowledge sources, most relationships (and the strongest of these) are with academia and are external to the Science Park in nature. In contrast to the pre-entrepreneurial phase, more firms report knowledge sources from the external relations in the financial sphere. More firms show stronger relations with policy/regulators and these interactions are concentrated in the technical and organisational spheres at all levels of proximity. Funding sources are external and in the financial sphere but some firms have acquired financial resources locally, in the Science Park, or through their personal network. Industrial partners are found at all levels, including the Science Park, and in all spheres; commercial partners are found outside the Science Park and the commercial sphere appears to produce few resources.

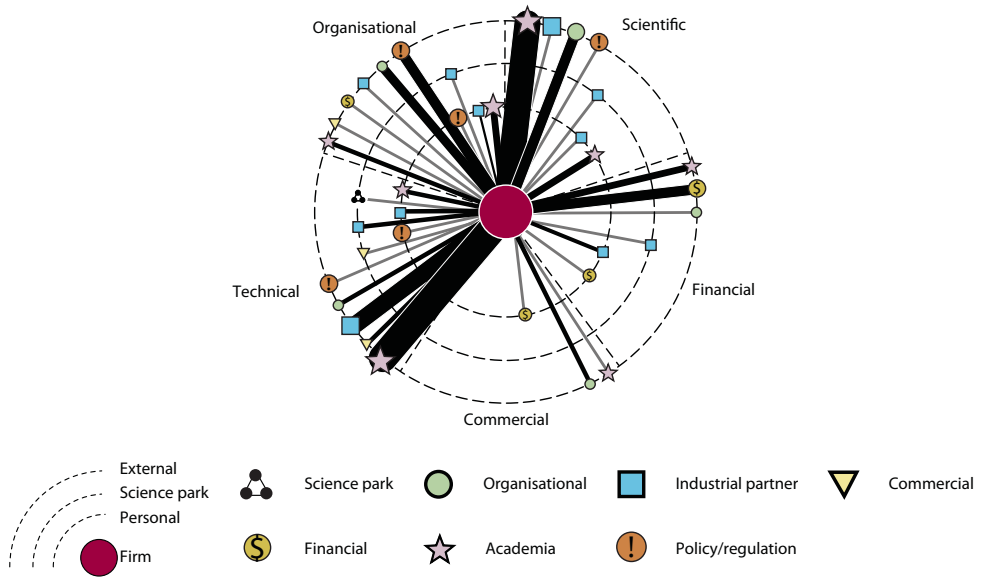
In terms of proximity, personal relationships are less significant; the occasional firm mentions funding or industrial partners and the most pronounced are relationships with academia. The role of the Science Park is more pronounced and diverse than in the pre-entrepreneurial phase, especially in the technical sphere. The Science Park itself appears as an actor in the organisational sphere and its main function seems to be to provide access to industrial partners. External relations are to knowledge sources in the scientific and technical spheres, and to industrial partners in the scientific, technical, and financial spheres. There are also, to a limited extent, emerging commercial relations.

The overall intensity of interactions has increased in this period, as compared to the pre-entrepreneurial phase. The overall number of firms reporting academic links in the scientific sphere also increases, with 8 companies compared to 7 over the previous period. The number of firms reporting interactions with venture capital (6) decreased from the previous period (9). This may be due to the firms becoming more financially sustainable as well as increasing efficiency in the use of their initial grant monies.

Managerial phase

Figure 3 shows a further decrease in the number of interactions in the financial sphere. Interactions with the Science Park administration were again limited to one firm, but within the technical sphere as opposed to the organisational sphere in the entrepreneurial phase. In this instance, the Science Park facilitated a technical exchange between firms on the Science Park. Academic interactions with industry increased with each phase, as reported by an increasing number of firms per phase. The number of firms reporting personal interactions remains relatively stable across the phases.

Figure 3 Leiden collective interactions during managerial phase



Rathenau Instituut

Note: Size of node indicates average number of interactions per firm. Edge thickness signifies count of firms reporting interactions. Grey edges signify only one firm reporting interaction.

Relations with policy-makers/regulators are stronger than in preceding phases. Funding has become gradually less important in the networks of founders - mostly in the financial sphere and external - and the financial sphere is less prominent than before. In relation to network partners, the Science Park mainly serves to find industrial partners and these are found in all spheres, except the commercial sphere, but especially in the technical sphere. However, the intensity and count of interactions in the commercial sphere are low and sparse.

In terms of proximity, personal relations primarily draw on knowledge, industrial partners and occasionally, funding opportunities. Relations/contacts within the Science Park draw in industrial partners and little else, and as such the Science Park now has a role in the technical sphere. External interactions result in knowledge from the scientific and technical spheres, industrial partners and organisational relations.

Summary
Overall interactions are primarily external, with sustained levels of interactions within the scientific and technical spheres. On review of the interview data, this increase is due to the development of relationships with industrial partners and many firm founders cited an increased feedback between themselves and the manufacturers of their products. This feedback led to the firms' improvement of their research and scientific practices.

Interactions mediated by the Science Park or with the Science Park administration directly, were minimal with only 1 firm reporting any significant interactions. Interactions with other firms within the Science Park were also minimal, with the few interactions being between the firm founders and their former academic supervisors, although many firms indicated that they would like to form commercial or scientific relationships with other firms in the Science Park. A reason cited by a firm founder - for their lack of interactions with other firms on the Science Park - was that either their services or products were not necessary to other firms, or other firms' services or products were, in return, not needed.

The external interactions reported by the founders were overwhelmingly international. Whilst all firms indicated they aimed to service the regional or national market, very few had found viable (in the firms' eyes) customers in the Netherlands.

The founders commonly reported interactions with universities or public research institutes. However, these interactions were the result of many of the firm founders serving as active faculty members within the universities named. Firm founders who were active faculty members whilst operating in the firm maintained a separation between the work they conducted at the firm and their research conducted in the university.

5.5.2 Scientific and technological output

Technological and scientific relevance
Figure 4 illustrates the linkages between the technologies (patent applications) and scientific output (publications) of one of the firms in our set. It shows the research streams of this example's founder, mapping out the evolution of research topics. The non-patent literature references cited by the patent applications have been included in the founder publication corpora. The example firm's founder is a prolific publisher, covering multiple research streams.

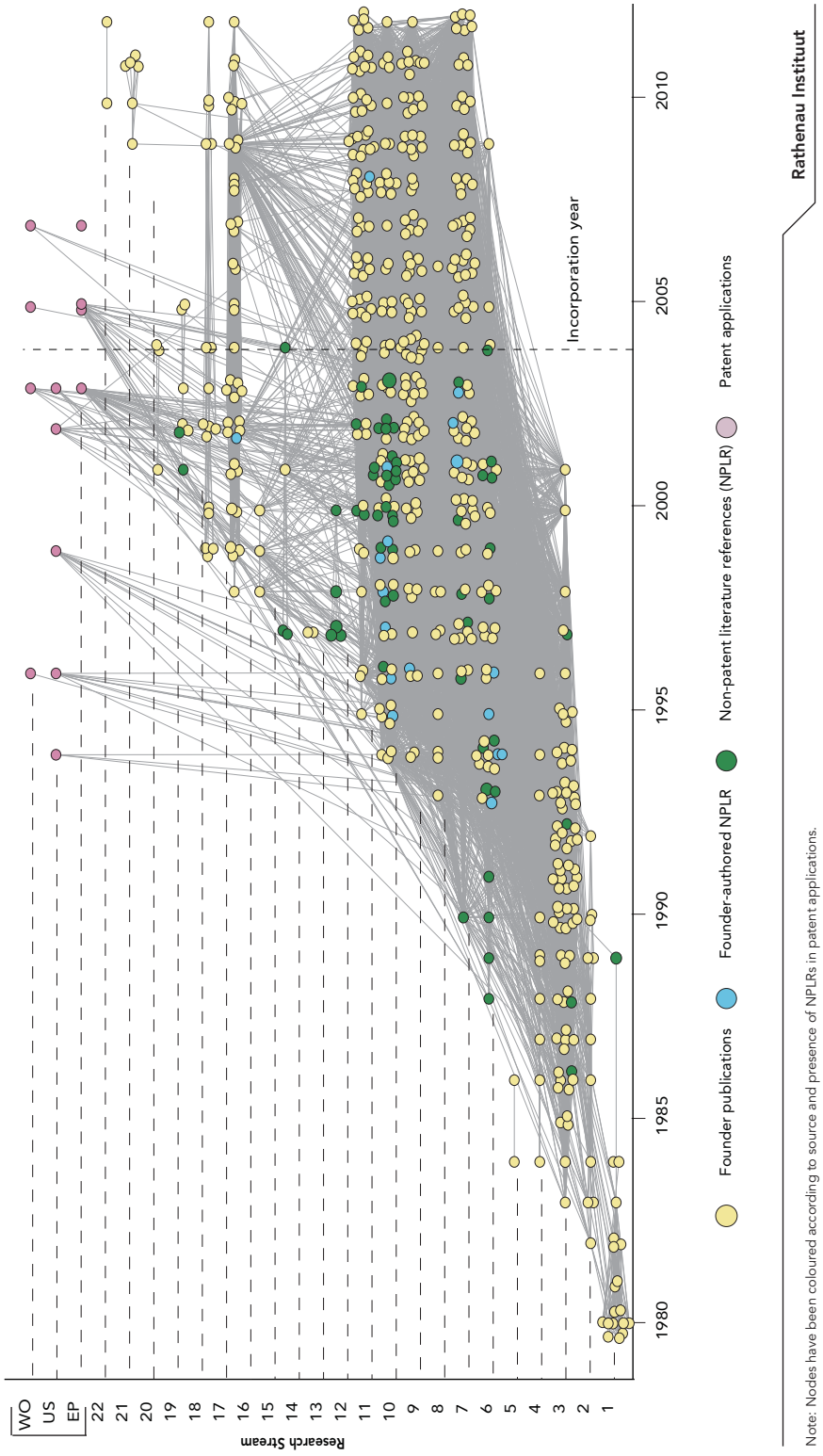


Figure 4 Example firm founder publications, patent applications and NPLRs.

Note: Nodes have been coloured according to source and presence of NPLRs in patent applications.

The founder applied for patents prior to firm formation, and in different patent authorities. The founder's research may be considered directly relevant to the technologies patented as the references cited by the patent applications are clustered together with the founder's publications and in many cases cite the founder's publications. The NPLRs precede major research streams, suggesting that during the preparation of the technologies embodied by the patent applications, the firm founder recognised and developed the necessary research skills and content to develop further their technologies.

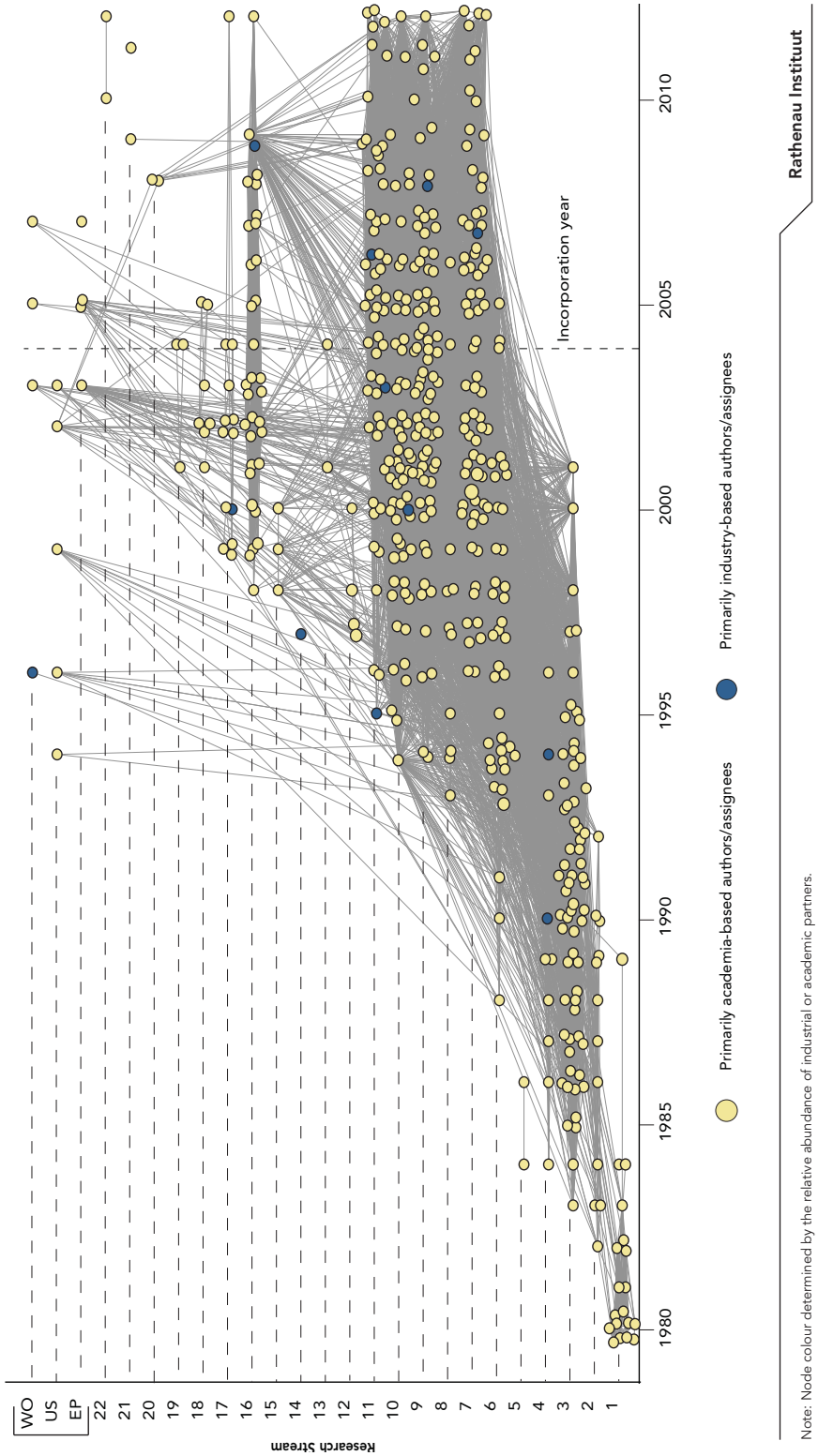
All active research streams prior to formation remain active, suggesting continuity of research involving the founder. There were, however, only two new streams of research by the founder after formation where the firm has converged on what it considers the most viable research stream for its technologies.

Table 1 gives a summarised account of the research streams in all nine firms in the set. The publishing propensity and diversity differs by firm founder - as can be seen by the number of research streams, and the number of active streams in the incorporation year of the firm. The number of new streams of research after incorporation is minimal. However, two firm founders have increased the number of streams compared to the number of active streams at incorporation. The continuity of research involving the founder is low for all firms except one, as seen by the number of streams the founders were involved with in 2011. The number of active streams with NPLRs present in the stream, indicating a strong link with the patent applications is high for all founders with active research streams at incorporation.

Table 1 Founder publishing research streams

	Number of Streams						
FIRM ID	Total	Active at incorporation	New after incorporation	Active 1 year after incorporation	Active 3 years after incorporation	Active in 2011	Active at incorporation with NPLR
1	9	5	0	5	3	1	5
2	20	10	2	9	7	8	7
3	7	1	1	1	0	0	0
4	7	3	1	3	3	1	3
5	14	3	1	2	0	0	0
6	4	0	2	0	0	1	0
7	2	1	1	1	1	2	1
8	3	0	0	0	0	0	0
9	10	1	2	1	1	3	0

Figure 5 Example firm founder publishing and patenting activities with industrial versus academic collaborators



Note: Node colour determined by the relative abundance of industrial or academic partners.

Academic and industrial collaborations

Figure 5 shows the publishing and patenting collaborations between the example firm founder, and academia and/or industry. The nodes are coloured according the degree of involvement of academic or industrial authors (publications) and assignees (patent applications).
There has been, in comparison to the total number of founder publications, little collaboration with industrial partners. Prior to firm formation, repeated industrial collaborations with different industrial partners occurred frequently. After incorporation, there was only one repeat industrial collaborator. However, the number of unique industrial collaborators *per publication* increased, with, for instance, one publication featuring three unique industrial partners.

The primary assignees on the example firm's patent applications are the firm founder's home university and public health laboratories. There are only two industrial assignees, and they are both on the same one application prior to firm formation. There are no patent applications with the firm as assignee.

Tables 2 and 3 present the results of the full set of firms. Table 2 shows that for most of the firms with active research streams at incorporation, the authors primarily come from academia. Firms 3 and 9 have a large portion of their authors from industry, with Firm 9 showing a varied mix of academic and/or industrial authors. This may be due to the fact that the products on offer by Firm 3 require less regulatory supervision as the product "[…] cannot be tested on humans[…]" and animal testing is the only solution, and "[..] the product either works or doesn't." In the case of Firm 9, the development of the hardware utilised in their services, was conducted in conjunction with theoretical advances from academia and engineering advances from collaboration with their technical partners.

Table 2 Academic and industrial collaboration composition (%) of founder publishing research streams active at incorporation

FIRM ID	Academic	Predominantly Academic	Academic and Industrial	Predominantly Industrial	Industrial
1	89	1	7	1	2
2	98	1	1	0	0
3	67	0	0	33	0
4	99	0	1	0	0
5	93	0	7	0	0
6	-	-	-	-	-
7	82	9	9	0	0
8	-	-	-	-	-
9	30	10	40	20	0

Table 3 presents the assignee composition of the patenting efforts of all the firms. Firms 3, 7 and 8 have industrial assignees exclusively, and Firm 9 has a vast majority of industrial assignees (94%). Firm 1 patents almost equally with academic and industrial assignees, whilst the rest of the firms tend to academic assignees. This again suggests that the products or services on offer from the firms are developed within a more technical environment, such as in the production methods of the product, or machinery required for the service.

Table 3 Academic and industrial collaboration composition of patent assignees pre- and post-incorporation

		Assignee origin									
		Academic		Predominantly Academic		Academic and Industrial		Predominantly Industrial		Industrial	
Firm ID	Total application count	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
1	67	16	14	0	0	19	0	3	2	9	4
2	13	7	5	0	0	0	0	1	0	0	0
3	6	0	0	0	0	0	0	0	0	4	2
4	16	8	3	0	0	0	0	0	0	2	3
5	7	-	0	-	0	-	0	-	0	-	7
6	30	-	0	-	0	-	9	-	1	-	20
7	10	0	0	0	0	0	0	0	0	2	8
8	9	0	-	0	-	0	-	0	-	9	-
9	45	3	0	0	0	0	0	0	0	34	8

Rathenau Instituut

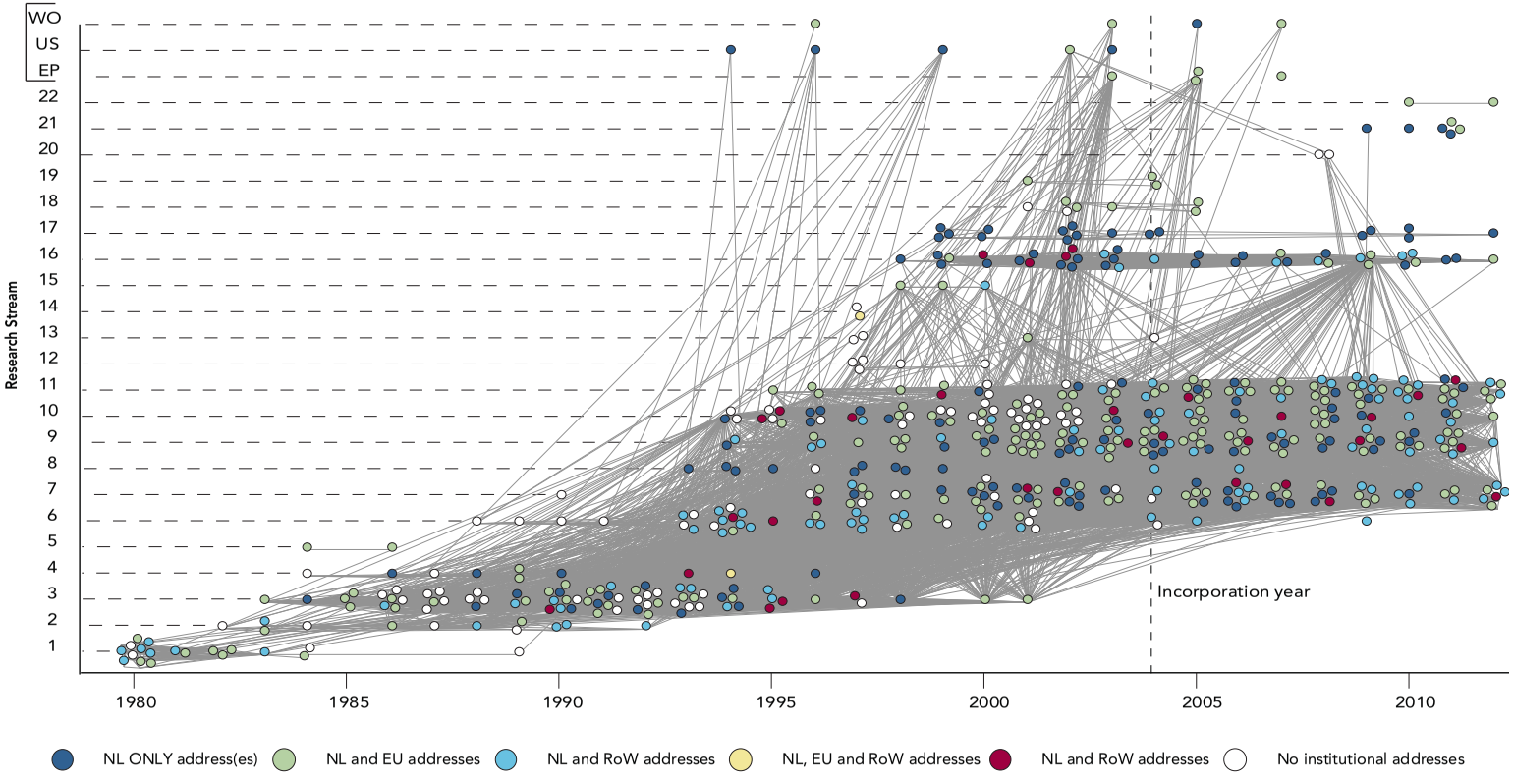
Local, regional and international academic collaborations

For our example firm, the number of publications which only have addresses in the Netherlands is consistent in all research streams still active after firm formation. The numbers of collaborations with EU addresses & NL addresses, and those with NL addresses & EU addresses & Rest of the World (RoW) addresses also remain consistent across research streams. There is no significant increase in the degree of academic internationalism after firm formation. Considering the scientific impact of the example firm, an increase in international collaboration from before the firm's formation to afterwards indicates that the firm has retained its international scientific reputation. The firm benefits from the variety and exposure of publishing collaborations.

The patenting activity of our example firm shown in Figure 6 has no EU academic partners (and only one Dutch academic assignee). There are multiple RoW partners, both before and after incorporation.

Tables 4 and 5 show the academic collaborations of all the firms in our set. Table 4 indicates the geographic distribution of academic collaborators in the publishing streams of the firms. All the firms publish with at least one academic address in the Netherlands. Firm 7 publishes exclusively

Figure 6 Example firm founder publishing and patenting activities with academic collaborators



Note: Node colours are determined by academic author addresses.

Rathenau Instituut

with Dutch academic partners and Firms 1 and 8 publish extensively with only Dutch partners. Firm 5 is more international in that it publishes almost exclusively with Dutch and RoW academic partners.

Table 4 International composition (%) of academic collaborators of founder publishing research streams active at incorporation

FIRM ID	NL only	NL & EU	NL & RoW	NL & EU & RoW
1	71.3	14.3	6.4	8.1
2	36.2	21.5	36.9	5.5
3	50	16.7	33.3	0
4	31.7	57.1	4.1	7.2
5	0	0	95.4	4.6
6	-	-	-	-
7	100	0	0	0
8	-	-	-	-
9	80	0	10	10

Rathenau Instituut

Links with Leiden University are seen as an integral part of the Science Park, and are reported as such by all the firms in the set. Table 5 shows the composition of academic collaborations with Leiden University, and with other universities and knowledge institutes in the Netherlands.

Table 5 Composition (%) of Dutch academic collaborators of founders

	Active at incorporation			New after incorporation			Active 1 year after incorporations			Active 3 years after incorporation			Active in 2011		
FIRM ID	Leiden	Other Uni	KI	Leiden	Other Uni	KI	Leiden	Other Uni	KI	Leiden	Other Uni	KI	Leiden	Other Uni	KI
1	81.3	7.3	3.8	-	-	-	81.3	7.3	3.8	81.3	9.8	3.6	75.9	12	0
2	22	23.5	11.2	100	100	0	22	23.5	3.9	9	20.3	3.9	23.1	22.2	3.9
3	0	20	80	0	0	0		20	80	-	-	-	-	-	-
4	95.5	2.7	2.7	66.6	16.6	0	95.5	27	2.7	95.5	2.7	2.7	86.5	2.7	2.7
5	0	0	0	0	0	0	0	0	0	-	-	-	-	-	-
6	-	-	-	75	62.5	0	-	-	-	-	-	-	75	25	0
7	76.9	7.7	7.7	62.5	18.7	0	76.9	7.7	7.7	76.9	7.7	7.7	69.7	15	7.7
8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	90.9	9.1	0	55.7	27.1	20	90.9	9.1	0	90.9	9.1	0	67.4	21.1	20

Rathenau Instituut

Note: KI - Knowledge Institutes

The Netherlands is a geographically compact nation, meaning that firms that are located at the Science Park are not limited to the closest universities for recruiting scientists. However, for firms with active research streams at incorporation, the vast majority of academic collaborations are with Leiden University. This is to be expected as many of the firm founders maintain active faculty positions at Leiden. There are also large numbers of collaborations with other Dutch universities, such as Firm 5, which had only published with other universities and knowledge institutes in the Netherlands prior to incorporation. Upon, and after, incorporation, all of Firm 5's publishing collaborations were with universities in other countries. Firm 3 published most frequently with knowledge institutes rather than universities. All new research streams developed by the firms after incorporation involved Leiden University, and all streams active in 2011 feature a majority of participants coming from Leiden University.

For all firms, the vast majority of their academic patent collaborators are from the Netherlands except for Firm 2 who collaborates with RoW academic partners on 30.8% of their pre-incorporation applications. Firm 5 collaborates with EU academic partners on 27.3% of their post-incorporation applications. For firms 3, 7 and 8, there are no academic collaborations at all either before or after incorporation.

Industrial collaborations

For the example firm, Figure 6 shows there was little industrial collaboration to speak of, with most collaborators being academic in nature. There are no patenting or publishing collaborations with any other firms within the Leiden BioScience Park nor any publications involving only Dutch industrial collaborators. There are also minimal collaborations with EU and RoW industrial partners. Whilst the minimal number of industrial partners may not be important, the lack of any *Dutch* industrial partners is. This suggests that scientific research conducted by the firm founder may have relevance to academia locally or elsewhere in the country (as seen in Results Figure 6), but little relevance for Dutch industry.

Table 6 shows the international composition of the industrial assignees of the patent applications of the firms, before and after incorporation. Half the firms began to develop their patent stock prior to incorporation, and 3 of the firms developed over 80% of their knowledge stock after incorporation. All but 3 of the firms have only Dutch and/or EU industrial assignees. The patent stocks of the 3 firms with Rest of World (RoW) industrial partners were developed prior to incorporation. Significantly, there are no other firms from the Leiden BioScience Park listed as assignees in any of the firms' patent applications.

Table 6 International co-assignees composition (%) of total assignees pre- and post-incorporation

	NL only		NL & EU		NL & RoW		NL & EU & RoW	
Firm ID	Pre	Post	Pre	Post	Pre	Post	Pre	Post
1	62.2	13.5	18.9	2.7	2.7	0	0	0
2	0	-	0	-	0	-	100	-
3	0	28.6	28.6	0	42.9	0	0	0
4	40	40	0	20	0	0	0	0
5[a]	-	100	-	0	-	0	-	0
6[a]	-	100	-	0	-	0	-	0
7	0	80	20	0	0	0	0	0
8	5.3	52.6	42.1	0	0	0	0	0
9	81	19	0	0	0	0	0	0

Note: a All applications are with firm as only industrial assignee.

Rathenau Instituut

5.6 Conclusions and discussions

We have examined in detail the social interactions and knowledge output of the founders of nine firms at Leiden BioScience Park, focusing on the following issues: (i) knowledge and knowledge capture, (ii) bridging social capital, (iii) the relationship between the two, and (iv) the role of the Science Park in facilitating access to social capital and aiding knowledge capture.

In terms of (i) knowledge and knowledge capture, the firm founder begins, in most cases, as an academic researcher. The research processes and decisions in academia are governed by specific scientific and social search strategies. These strategies shape the eventual quanta of knowledge that go on to become a technology. Once this knowledge has been recognised as being exploitable, the idea for a commercial exploitation route is formed. Having recognised the exploitability of a particular research stream or area, the motivations, rules and norms of research in academia change, by necessity, to accommodate the increasing number and variety of stakeholders, as well as the supporting infrastructure found in a Science Park.

The firms were still employing the scientific research that they had conducted and the expertise that they had developed to their incorporation. Only one founder of the nine who had active research streams at incorporation did not publish after incorporation. Of those that did publish after incorporation, all but one founder engaged in new research streams after incorporation. This would imply that there are still active exploration efforts involving the founders at the firms. There is continuity of the founders' research in all but 2 of the firms, with active streams at incorporation mostly still active at least 3 years after incorporation.

March's (1991) exploration and exploitation strategies were evident, as seen in the breakdown of publishing and patenting activities of the firms. Most firms struck balances between explorative activities and exploitative activities in punctuated equilibrium (Gupta et al. 2006).

In terms of (ii) bridging social capital, the interactions of the firm founders before and after the incorporation of the firm demonstrate a clear trend in that they are primarily external to the Science Park. There were no significant scientific or technological interactions (and associated private benefits) reported by the firms with any other firms on the Science Park. One founder reported an interaction with another firm, but this was between the founder and his previous employer who is also located on the Science Park. There are significant scientific and technological interactions with industrial partners, with the interviewed firm founders stating that these inter-actions play a large role in the development of their respective technologies. For founders reporting international collaborations with industry, the primary reason stated is that the products or services they offer do not have any regional or national relevance. However, the interviewees in these cases stated that they would like to have commercial relations with customers in the same region and country.

In terms of (iii) the combination of bridging social capital and knowledge creation and capture, all nine firms have patent applications that are related to the research conducted by the founder prior to applying for the patent(s). The academic and industrial composition of the patents reflected much of what was reported by the founders during the interviews. For those founders that were still serving as active faculty members at universities, the assignees of the patents were primarily their host universities. After incorporation, the presence of industrial assignees increased significantly. For the founders with no patent applications before incorporation, most of the assignees were industrial partners, with few universities listed as assignees. This suggests that prior to firm incorporation, the university plays the largest role in the technologies, but after incorporation there is increased interest from industry in the services and products.

For the firms interviewed, there appears to be an imbalance in the research they conduct in terms of their collaborators. Their scientific publications prior to incorporation strongly feature international partners. After incorporation, only firms with founders still active at the university continue a high level of international academic collaboration. For firm founders who are not active faculty members, publishing activities become increasingly local, mostly with their alma mater and increasingly with the university affiliated with the Science Park, in this case Leiden University. The share of international collaborators in the scientific output of the founders is high to begin with for most of the firms. This could be considered an artefact of the founders' previous collaborations in a university setting where many of the collaborators were from universities in other countries. Although the number of international *industrial* collaborations varies across the firms, both in patenting and publishing, there seems to be no significant change to the internationalism of industrial partnerships in research. At the same time, most the founders interviewed found that customers for the products and services they provide are not found locally or even in the same region. Rather, they are mostly from elsewhere in the EU or further afield. This would suggest that whilst there may not be a local market for their products and services, the research conducted towards developing the products and services is strongly helped by the local education and industrial sectors.

In terms of (iv) the role of the Science Park in facilitating access to social capital and knowledge capture opportunities, the sum of the reported interactions runs contrary to many of the stated goals of a Science Park. The most notable of these is that there be interaction between firms within the Science Park and its administration, so that they exploit the network benefits of locating to a Science Park. The principal scientific and technological pull of the Science Park as reported by the founders was the proximity to the local HEI, a motivation also reported in Löfsten & Lindelöf (2003).

We believe that the level of detail in our study outweighs the restrictive selection criteria. We add a new dimension to future studies on Science Parks, and academic entrepreneurs who choose to locate to Science Parks. We feel that previous research has too often neglected the core components of a Science Park, and the firms located there. That is to say the effect on the technologies and research processes of the founders and their ability to mobilise social capital. The quantitative aspect of our study can provide insight in further studies for policy-makers as to the historical development and level of collaboration between firms located on Science Parks and the internationalism of academic or industrial collaborations. Our qualitative approach can be of help to policy-makers when re-examining the purported benefits of a Science Park and whether a Science Park is in fact the ideal carrier for these benefits.

The results of this paper add new weight to the need for a careful re-examination of the role of the Science Park in regional and national policy discussions. As the Netherlands is a geographically compact country, many of the logistical benefits may be moot. However, our results seem to suggest that the close association of start-ups with the local university and with national industrial partners lead to a more productive firm, both in terms of new research and associated output.

5.7 References

Adler, P.S. & Kwon, S.-W. (2002). Social Capital: Prospects for a New Concept. *Academy of Management Review, 27*(1), 17-40.

Audretsch, D.B., Lehmann, E. E. & Warning, S. (2005). University spillovers and new firm location. *Research Policy, 34*(7), 1113-1122.

Barney, J.B. (2001). Resource-based theories of competitive advantage: A ten-year retrospective on the resource-based view. *Journal of management* 27(6): 643-650.

Blondel, V.D., Guillaume, J.L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment,* 2008, P10008.

Bozeman, B., Dietz, J.S. & Gaughan, M. (2001). Scientific and technical human capital: an alternative model for research evaluation. *International Journal of Technology Management, 22*(7), 716-740.

Cainelli, G., Mancinelli, S. & Mazzanti, M. (2007). Social capital and innovation dynamics in district-based local systems. *Journal of Socio-Economics, 36*(6), 932-948.

Castells, M. & Hall, P. (1994). *Technopoles of the World: The Making of 21st Century Industrial Complexes*: Routledge.

Clarysse, B., Wright, M., Lockett, A., van de Velde, E. & Vohora, A. (2005). Spinning out new ventures: a typology of incubation strategies from European research institutions. *Journal of Business Venturing, 20*(2), 183-216.

Costas, R., van Leeuwen, T.N. & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology, 61*(8), 1564-1581.

Das, T.K. & Teng, B.S. (1997). Time and Entrepreneurial Risk Behavior. *Entrepreneurship Theory and Practice, 22*(2), 69-71.

Das, T.K. & Teng, B.S. (2000). A resource-based theory of strategic alliances. *Journal of management, 26*(1), 31-61.

De Carolis, D.M. & Saparito, P. (2006). Social capital, cognition, and entrepreneurial opportunities: A theoretical framework. *Entrepreneurship Theory and Practice, 30*(1), 41-56.

Deeds, D.L. & Hill, C.W.L. (1996). Strategic alliances and the rate of new product development: An empirical study of entrepreneurial biotechnology firms. *Journal of Business Venturing, 11*(1), 41-55.

Dettwiler, P., Lindelöf, P. & Löfsten, H. (2006). Utility of location: A comparative survey between small new technology-based firms located on and off Science Parks--Implications for facilities management. *Technovation, 26*(4), 506-517.

Dubine, P. & Aldrich, H. (1991). Personal and Extended Networks are Central to the Entrepreneurial Process. *Journal of Business Venturing, 6*(5), 305-313.

Elfring, T. & Hulsink, W. (2003). Networks in entrepreneurship: The case of high-technology firms. *Small business economics, 21*(4), 409-422.

Felsenstein, D. (1994). University-related science parks -- 'seedbeds' or 'enclaves' of innovation? *Technovation, 14*(2), 93-110.

Fukugawa, N. (2006). Science parks in Japan and their value-added contributions to new technology-based firms. *International Journal of Industrial Organization, 24*(2), 381-400.

Greve, A. & Salaff, J.W. (2001). The development of corporate social capital in complex innova tion processes. In S. M. Gabbay & R. Leenders (Eds.), *Social Capital of Organizations* (Vol. 18, pp. 107-134): Emerald Group Publishing Limited.

Gupta, A.K., Smith, K.G. & Shalley, C.E. (2006). The interplay between exploration and exploitation. *Academy of Management Journal, 49*(4), 693-706.

Gurney, T., Schoen, A., Horlings, E., Sumikura, K., Laurens, P., van den Besselaar, P. et al. (2012). *Knowledge Capture Mechanisms in Bioventure Corporations.* Paper presented at the 17th International Conference on Science and Technology Indicators (STI), Montreal.

Hagstrom, W.O. (1974). Competition in science. *American sociological review*, 1-18.

Ho, M.W.Y. & Wilson, M. (2007). Biotechnology founders and employment systems of start-ups. *International Journal of Technology, Policy and Management, 7*(3), 263-279.

Horlings, E. & Gurney, T. (2012). Search strategies along the academic lifecycle. *Scientometrics*, 1-24.

Koh, F.C.C., Koh, W.T.H. & Tschang, F.T. (2005). An analytical framework for science parks and technology districts with an application to Singapore. *Journal of Business Venturing, 20*(2), 217-239.

Lanciano-Morandat, C., Jolivet, E., Gurney, T., Nohara, H., van den Besselaar, P. & Pardo, D. (2009). Le capital social des entrepreneurs comme indice de l'émergence de clusters ? *Revue d'économie industrielle*(4), 177-205.

Lancichinetti, A. & Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical Review E, 80*(5), 056117.

Landry, R., Amara, N. & Lamari, M. (2002). Does social capital determine innovation? To what extent? *Technological Forecasting and Social Change, 69*(7), 681-701.

Leana, C.R., & van Buren, H.J. (1999). Organizational social capital and employment practices. *Academy of management review, 24*(3), 538-555.

Levin, S.G. & Stephan, P.E. (1991). Research productivity over the life cycle: evidence for academic scientists. *The American Economic Review*, 114-132.

Lin, N. (1999). Building a network theory of social capital. *Connections, 22*(1), 28-51.

Link, A.N. & Scott, J.T. (2007). The economics of university research parks. *Oxford Review of Economic Policy, 23*(4), 661-674.

Löfsten, H. & Lindelöf, P. (2003). Determinants for an entrepreneurial milieu: Science Parks and business policy in growing firms. *Technovation, 23*(1), 51-64.

Löfsten, H. & Lindelöf, P. (2005). R&D networks and product innovation patterns--academic and non-academic new technology-based firms on Science Parks. *Technovation, 25*(9), 1025-1037..

March, J.G. (1991). Exploration and exploitation in organizational learning. *Organization science, 2*(1), 71-87.

Merton, R.K. (1957). Priorities in scientific discovery: a chapter in the sociology of science. *American sociological review, 22*(6), 635-659.

Merton, R.K. (1969). Behavior patterns of scientists. *American scientist, 57*(1), 1.

Murray, F. (2004). The role of academic inventors in entrepreneurial firms: sharing the laboratory life. *Research Policy, 33*(4), 643-659.

Oliver, A.L. (2004). Biotechnology entrepreneurial scientists and their collaborations. *Research Policy, 33*(4), 583-597.

Palla, G., Derényi, I., Farkas, I. & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature, 435*(7043), 814-818.

Parise, S. & Henderson, J.C. (2001). Knowledge resource exchange in strategic alliances. *IBM Systems Journal 40*(4), 908-924.

Phillimore, J. (1999). Beyond the linear view of innovation in science park evaluation An analysis of Western Australian Technology Park. *Technovation, 19*(11), 673-680.

Quintas, P., Wield, D. & Massey, D. (1992). Academic-industry links and innovation: questioning the science park model. *Technovation, 12*(3), 161-175.

Rothaermel, F.T. & Deeds, D.L. (2004). Exploration and exploitation alliances in biotechnology: A system of new product development. *Strategic management journal, 25*(3), 201-221.

Schartinger, D., Rammer, C., Fischer, M.M. & Fröhlich, J. (2002). Knowledge interactions between universities and industry in Austria: sectoral patterns and determinants. *Research Policy, 31*(3), 303-328.

Shearmur, R. & Doloreux, D. (2000). Science parks: actors or reactors? Canadian science parks in their urban context. *Environment and Planning A, 32*(6), 1065-1082.

Siegel, D.S., Westhead, P. & Wright, M. (2003). Assessing the impact of university science parks on research productivity: exploratory firm-level evidence from the United Kingdom. *International Journal of Industrial Organization, 21*(9), 1357-1369.

Somers, A., Gurney, T., Horlings, E., & van den Besselaar, P. (2009). *Science Assessment Integrated Network Toolkit (SAINT): A scientometric toolbox for analyzing knowledge dynamics*. The Hague: Rathenau Institute.

Sorenson, O. (2003). Social networks and industrial geography. *Journal of Evolutionary Economics, 13*(5), 513-527.

Squicciarini, M. (2008). Science Parks' tenants versus out-of-Park firms: who innovates more? A duration model *The Journal of Technology Transfer, 33*(1), 45-71.

van den Besselaar, P. & Heimeriks, G. (2006). Mapping research topics using word-reference co-occurrences: a method and an exploratory case study. *Scientometrics, 68*(3).

Westhead, P. & Batstone, S. (1998). Independent Technology-based Firms: The Perceived Benefits of a Science Park Location. *Urban Studies, 35*(12), 2197-2219.

Zucker, L.G. & Darby, M.R. (1996). Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry. *Proceedings of the National Academy of Sciences of the United States of America, 93*(23), 12709.

Zuckerman, H. (1992). The proliferation of prizes: Nobel complements and Nobel surrogates in the reward system of science. *Theoretical Medicine and Bioethics, 13*(2), 217-231.

Zuckerman, H. & Cole, J.R. (1994). Research strategies in science: A preliminary inquiry. *Creativity Research Journal, 7*(3-4), 391-405.

6 Conclusions

This dissertation has examined the routes, processes and environments of knowledge production and transfer. At first glance, the overall route from knowledge production in a lab setting to the social interactions amongst firm founders located within a Science Park may appear convoluted.

The primary question of this dissertation (*what knowledge elements are transferred from academia to industry, how are they transferred, and what factors influence this transfer?)* draws upon larger theories of knowledge transfer in general and the effect of the environment, be it social or physical, on innovation. In order to answer the research question, we have to identify the researchers performing the process, the knowledge elements involved and the conditions under which knowledge transfer operates. This leads to the following sub-questions: (i) *how can we disambiguate researchers with an effective balance between precision and recall?; (ii) how can we identify knowledge elements and their attributes in an operational way, and what elements are transferred between actors?*; and *(iii) what resources, and from which actors and operational spheres, contribute most significantly to the development of an academic spin-off and its host technology?*

In this final chapter, I synthesise the conclusions of the four studies and how they relate to the primary question and sub-questions. These findings add to the existing knowledge and theoretical bases underpinning my research. The theoretical implications of the results and conclusions will be examined, with special regard to further applications.

6.1 How can we disambiguate researchers with an effective balance between precision and recall?

Scientometric studies have become increasingly important in the evaluation and analysis of research systems. Key to these evaluations and analyses is the availability of clean input data, and efficient and balanced disambiguation techniques are vital. We developed a method that relies on using the best possible combination of metadata available to us, based on the pitfalls of current techniques.

Three problems are commonly found with current techniques: discarding of data, limited selection of metadata, and a lack of consideration for evolving research streams and actual contributions to research by authors. Our solutions to handling the first two problems, data discarding and limited metadata, are interrelated: we selected the best possible alternate combination of available metadata. In other words, for records missing one or more fields, the proportional discriminating powers of the available metadata were adjusted depending on available combinations. All this occurs 'on the fly' within the algorithm, based on previous calibrations. The last problems, those of the evolution of terminologies (Healey et al., 1986), research streams or knowledge homogeneity (Tang & Walsh, 2010) and author contributions (Bates et al., 2004; Moed, 2000; Yank & Rennie, 1999), rely on recognising the changing roles, topics and requirements of researchers as they progress through their academic life cycle. For department heads, author status is occasionally 'honorary', in that much of the research and write-up was actually performed by others. In evaluations, however, every author listing counts. In other current techniques, similarity calculations assume input from *all* authors, in *equal* parts. Relying on this

assumption leads to networks of publications by authors, exhibiting low or no similarity. The algorithms developed in Chapter 2 tackle this problem by assuming that contributions by authors vary according to listing (an exception is made for alphabetically ordered listings). The discriminating powers of the indicators were adjusted to allow for this. For example, the importance of the title or abstract words (most likely selected by the authors writing the paper) is reduced, whereas the journal name (an aspect more likely to be decided on strategic grounds (Leydesdorff et al., 1994) by the head of department) is granted more discriminatory power in the algorithm.

A similar practice is deployed to adjust for the evolution of research streams. Over time, a researcher's usage of words will change to reflect changing fields of interest and publishing. When comparing records published ten years apart, by apportioning less discriminatory power to title or abstract words but more power to journal field characterisation or cited references, we account for the evolution of researchers' interests. Taking the dynamic selection of alternate metadata together with the adjustments for evolving research interests and contributions means that the developed algorithm allows for more accurate and inclusive results. The work conducted in Chapter 2 both provides theoretical insight into the research and publishing practices of individuals and also contributes to the larger issue of disambiguation.

6.2 How can we identify knowledge elements and their attributes in an operational way, and what elements are transferred between actors?

In the third and fourth chapters of this thesis, we developed a novel method for the visualization and quantification of specific contributions, by individuals and institutions, to the developmental routes of specific quanta of knowledge. We then applied this method to an individual firm founder, looking at (a) how the scientific background of the patent corpus links to the scientific output of the inventor; (b) how a researcher operates in a collaborative environment, and if the contributions of those contributors are visible in the patent corpus; and (c) whether the inventor demonstrates a level of adaptive knowledge acquisition, necessary for the development of a technology.

To link patent and publication data, we used the non-patent literature references (NPLR) found in patent applications. We grouped together the NPLRs and the author/inventor's corpora of publications, linking their shared title words and cited references, and then they were clustered. The topical foreground and cognitive background of the research cited by the patent applications, and that of the author/inventor's whole corpus of publications, provided a clear view of the required knowledge platforms and absorptive capacity of the inventor and of the contribution it makes to the development and transfer of knowledge. Both the bibliographic (examiner) and in-text (applicant) NPLRs were included to provide better balance between what is considered important to the technologies (Karki, 1997; Tamada et al., 2006), in order to negate some of the effects of strategic citing in patent applications.

For further clarification, each of the inventor/author's research streams was differentiated further by the introduction of 'concept clusters'. With these concept clusters, we identified the specific, rather than general, contributions made by the researcher involved. Our approach took into account the role of co-authors and co-inventors, allowing us to determine the specific expertise these collaborators added to the technologies. The institutional affiliation of the inventors and authors also gave an institutional approach to the contributions to the technologies. Additionally,

the multidirectional aspect of knowledge and skill transfer between basic and applied research can be examined in minute detail using this method, including the absorptive capacity of individuals and institutions.

In relation to the research question of this dissertation, it is important to remember that technologies should be seen in the context of their surroundings, practices, artefacts and base understanding (Nelson, 2004). Absorptive capacity (Cohen & Levinthal, 1990) and the theoretical extensions of acquisition, assimilation, transformation and exploitation under potential and realised absorptive capacity (Zahra & George, 2002) are necessary to describe the roles of endogenous and exogenous sources of knowledge in the development of a technology. The transfer of knowledge between universities and start-ups is most often based upon a specific technology or scientific result (Baba et al., 2009; Carayannis et al., 1998). This can occur in various forms, such as technology or skills (Steffensen et al., 2000) or collaborations (Agrawal et al., 2006). They are typically either codified, such as publications and patents, or tacit, notably skill sets (Cohen et al., 2002). The role of a star or core scientist (Furukawa & Goto, 2006; Zucker & Darby, 1996) is important in facilitating this transfer.

To aid the analysis of specific contributions by individuals to a technology, several descriptors were introduced for (1) the reputational and applicability aspects of the scientific base work conducted by an individual; (2) the localisation of an individual's overall research trajectory in or outside the field(s) of research necessary for the technologies; (3) other fields of science being utilised by the technologies; (4) the level of input of collaborators; (5) the shared knowledge features (such as concepts, knowledge bases and, to a certain extent, skill sets) utilised by the technologies in relation to their sources; and (6) the individual incorporated skill sets acquired during the development of the technologies, and possibly applied to further basic research (knowledge creation feedback).

We found that during the initial stages of a technology's development, our individual in the case study, Professor Yusuke Nakamura, recognised the importance of exogenous knowledge sources. The expertise of his network of co-inventors and co-authors increased his scope and ability to source new knowledge required for the technologies. The technologies were initially based upon Nakamura's knowledge and skills set, but some aspects were outside his expertise. To combat this perceived gap in Nakamura's base knowledge, the assimilation of new knowledge was necessary, and this was ultimately visible in the patent and publication analysis output. Research cited by patent applications that did not fall within Nakamura's or his co-inventors' expertise was quickly incorporated into the research agenda. The result of this research was seen in the increased publication output of Nakamura in these sub-fields and in some cases the eventual citing of this supplementary research by the patent applications in subsequent developmental stages of the technologies.

6.3 What resources, and from which actors and operational spheres, contribute most significantly to the development of an academic spin-off and its host technology?

In the fifth chapter, we examined, in tandem, the social and scientific networks of the founders of biotechnology-orientated firms in the specific context of a Science Park. We were looking to

investigate the cognitive development routes of an idea generated in academia and exploited in industry, the relations supporting the knowledge capture and transformation, in particular the role and sources of the firm founders' social capital. Finally, we investigated the role of Science Parks in facilitating these processes.

Science Parks, and their utility, have been extensively studied, yet common definitions are hard to come by. General descriptions of a Science Park sum to a property-based, technology-orientated agglomeration of firms of varying specialisation and size, with close links and opportunities, either cognitive, geographical, structural or commercial, amongst firms and to a higher education or research institution (Das & Teng, 1997; Löfsten & Lindelöf, 2005; Quintas et al., 1992; Siegel et al., 2003). Each Science Park has unique origins - some were developed for the infrastructure, whereas other were developed to improve R&D innovation and production, or to provide intellectual development (Koh et al., 2005).

For firms choosing to locate within a Science Park, neoclassical location theory tends to dominate the decision processes of the firm founder (Westhead & Batstone, 1998). These typically include logistical issues, such as the proximity to the founder's home. This is not to say that firms interviewed considered only these issues, but rather that *practicalities* won over *potentialities.* In our interviews with firm founders, they all expressed interest in being able to collaborate with other firms in the Science Park (in line with one of the espoused benefits of locating within a Science Park). However, there was little evidence in the publication and patent data to show that they actually conducted collaborative research with other firms located at the Park.

That is not to say that there was no collaboration, rather there was no substantial evidence of collaboration. From the patent and publication data, the regional and international characteristics of co-assignees and co-authors show that for almost all of the firms, collaborative activities were common, but with firms *outside* the Science Park. Academic collaborations were primarily with the local HEI, Leiden University in this case, and a few founders maintained strong links with their alma maters beyond Leiden. This was reflected in the interview data where the interactions were internal, i.e. initiated before firm's formation and location to the Science Park, and external, i.e. with academic and industrial partners elsewhere in the country and abroad.

Social capital as a resource can be considered supplementary and enabling to the stock knowledge, financial capital and skills of an entrepreneur (Dubine & Aldrich, 1991; Greve & Salaff, 2001; Lin, 1999). In the interview data, for the firm founders who exploited their networks, in most of the spheres discussed in the chapter - particularly the scientific, technical and financial spheres - drew capital from either internal sources (i.e. historically through personal relationships prior to firm formation) or from sources external to the Science Park. Only a few firms reported any interactions of any nature with the Science Park administration or other firms located at the Science Park. Developing social capital through the discovery of opportunities, securing resources, and obtaining legitimacy (Elfring & Hulsink, 2003) means that the sources for these processes did not come from within the Science Park.

We found that the scientific capabilities of the firm founder were significant in developing and expanding the firm's scientific base, and for its eventual patent output. The substantial similarities

between the patent content and the scholarly output of the firm founder (our proxy being the co-location of NPLRs and the founder's publication corpora) and the number of active research streams at and after incorporation both suggest that the scientific base of the founder had a large supporting role.

Debates surrounding the utility of Science Parks are bound to continue. Future research is likely to feature arguments relating to what specific market failures a Science Park addresses (Siegel et al., 2003), with the supplied rationales ultimately undermined by inconsistencies in how Science Parks are defined. Their broad descriptions (examples of which are presented earlier in this section) encompass Science Parks of a range of sizes, with varying intensity of ties to HEIs, and with different administrative styles etc. Studies comparing Parks can, unfortunately, only be applicable to the Parks mentioned in their data sets. Additionally, each Science Park has unique origins and unique motivations for its formation. These again relate to the specific market failures being addressed.

The lifeblood of a Science Park is its tenants. Science Parks compete for tenants, and those tenants exist in a *coopetitive* environment (Nalebuff & Brandenburger, 1996). Tenants compete for access to networks and the benefits these networks bring. They also are presented with the opportunity to cooperate with other tenants in the park, sharing resources and mitigating risks whilst expanding their potential access to networks. To counteract the diversity issues in a Science Park in terms of evaluating their utility, we feel that more emphasis is needed on analysing the knowledge structures and competences of the firms within a Science Park.

6.4 All together

The overarching research question of this thesis (*what knowledge elements are transferred from academia to industry, how are they transferred, and what factors influence this transfer?)* is at first glance a broad question. It is necessarily broad so as to encompass the complexity of knowledge transfer in relation to absorptive capacity, social capital and the environment in which these knowledge transfer processes take place. In anticipation of the inevitable question of applicability, this thesis goes some way to providing a toolbox for parties that are interested in discovering which elements are transferred, where to and where from, and what factors influence this transfer, for *their* specific field of application. Each chapter in this thesis provides a methodological approach that can be applied to different cases, in different contexts. The case studies in this thesis are used to provide examples of how the methodologies should be applied, but also to validate their logic and results. As such, each chapter provides substantive examples of each element of the primary question with its own case study.

The underpinnings of this methodological toolbox begin with considering the effect of an individual's previous research on future research plans, and the similarities between past and present current research streams. The methods developed and insights provided in the chapter on disambiguation allow us to analyse factors such as the similarities and differences between research conducted during the PhD phase of a researcher's career and the professorial phase. Over time, an individual's research contributions may change with rank and eventual specialisation, but their incorporated knowledge and skill sets developed remain. Research conducted in academia and eventually applied in industry follows a convoluted path. We needed to gain an

understanding of this path for our disambiguation algorithms to succeed, and such an understanding provides a first glimpse at what knowledge elements are transferred over time.

The methodology and case study in the third and fourth chapters serve as a vehicle to examine the specific contributions of an existing knowledge base to the development of a technology platform, which involves identifying the knowledge elements and, to a certain extent, how they are transferred. The knowledge base does not necessarily come from one individual, but also from co-authors and co-inventors, and from other researchers working in different research settings. The methodology outlined in these chapters provided a toolkit for us to uncover the linkages between research conducted within academia and the eventual application of that research in industry, and the case study provided an example of what our approach can reveal. By applying our new method to a real case study, we demonstrated the ability to combine exogenously generated knowledge with a current knowledge base. New research conducted by Nakamura was guided by previous efforts, indicating that research practices and results are constantly evolving to inform, guide and provide the basis for extensions to different technologies. With detailed descriptions of the linked chains of research, we showed that a research corpus of an individual and their co-inventors and co-authors, can be readily recognised and identified in the exploitation of their research (i.e. in patents). The thematic links between the technologies and the underlying science was clearly identified in this chapter, verifying our methodological approach.

There is a long list of purported benefits for a firm to decide to locate to a Science Park. Its close proximity to a HEI and the prevalence of like-minded firms in the vicinity are mere examples of the reasons firms decide to locate to a Science Park. In social capital terms, a Science Park provides accessibility to resources, but for the firms in our case study, for the most part, these resources existed *in potentia*. It is important to note that for the firms under study there were no access barriers imposed by the Science Park. All the firm founders considered the Science Park to be an important *potential* source of collaborations and customers. If the opportunity arose, all the firm founders stated they would consider it. However, using the lens of only the purported benefits of locating to a Science Park, we failed to see more than practical benefits.

It is the firms that drive the success of a Science Park, and each firm is driven by its own scientific capacities and potential market linkages. For firms to truly enjoy the network benefits of a Science Park, there should be an overlap of not only their fundamental or applied sciences, but also of potential collaborators and customers. As such, we feel there needs to be a greater emphasis on the underlying sciences and technologies hosted by each firm when setting up a Science Park.

An important conclusion from this chapter was to view the primary research question from two perspectives: access to resources and technology development. We found it was necessary to blend the two perspectives, as many of the opportunities in one arise from the other. Ideas generated by a scientist in academia are not initially beholden to entrepreneurial dynamics within their network. They are, however, subject to the incentives structure of academia. New research streams are common in academia, where networks of scientists contribute to one another's research, iteratively guiding the development of an idea or technology. If a certain stream or idea

is deemed suitable for exploitation, the scientist/entrepreneur/founder's various networks simultaneously come into play, and the opportunities for further scientific development diminish. In their place, commercialisation of the idea becomes paramount.

Incentive structures shape both the strategies of academic researchers and industrial researchers in terms of valuing their research and results, and thus what aspects of the research will be developed and transferred. For academic spin-offs, access to scientific and technical networks and/or resources is not restricted to academia but extend into industry. The development of these resources is crucial to the development of the technology, but is secondary for the spin-off itself. It is access to the organisational, financial and regulatory networks that tops the priority list and begins to refine the technology.

Referring back to the quote by Leonardo Da Vinci at the beginning of this thesis, "the colour of the object illuminated partakes of the colour of that which illuminates it", this message becomes clearer when applied to knowledge transfer. The characteristics of an idea conceived in academia and transferred to industry vary in many ways but, using the developed tools, we can identify its origins, track its evolutionary path, and examine the effects of the environment.

6.5 Implications

From Chapter 2, in regards to future disambiguation, our approach confirms the need for an understanding of scientific research practices. Methods that employ purely statistical approaches often fail to accommodate the vagaries of research in practice. Each field has its own practices, which need to be taken into account in any disambiguation approach. For example, low citation rates, or a greater prevalence for single-author publications: both of these will affect an algorithm's discerning power.

A reliable and efficient disambiguation approach is extremely relevant to policy formation. For evaluation purposes, if corpora of publications are incomplete due to inaccurate disambiguation techniques, it can result in tremendous potential losses of funding for researchers. Furthermore, the rise of science systems from Asia, where there is a lesser degree of diversity amongst names, will make the issue of cleanly disambiguated data even more pressing.

There are other fields which have applications for disambiguation techniques, such as in social networks analysis and in search engine design. A future extension of our approach would incorporate heterogeneous data sets including professional social network data, and blog or self-published data, in addition to publication and patent data. The use of these 'alt-metrics' is growing in popularity as research - and science in general - is conducted at multiple levels of engagement and dissemination. The need for effective disambiguation techniques to cover the growing heterogeneity of data sources is absolutely required.

From Chapters 3 and 4, we believe that this method of mapping science to technology could deepen our understanding of the contributions made not only by individuals, but also by university departments and firms - providing us with data which can be used as input for theoretical systemic models such as those that concern funding instruments and policy. Technology positioning and evaluation models for funding allocation can be tested more completely by

examining the relative contributions of exogenous or endogenous sources of knowledge per field, allowing models to be adapted to suit varying publication and patenting output between fields. Policy-makers can be better informed on the adoption rates of indigenously-produced knowledge, with a deeper understanding of research competences in their own countries and how they compare to science systems elsewhere in the world.

In addition, the specificity of this approach is useful to all those involved in research. For universities, understanding specific contributions to a group of technologies can help TTOs to recognise what forthcoming research may be of value, either to their own IP portfolios or those of their industrial partners. This can similarly be useful on a regional or national level, with funding instruments gaining the ability to link and identify their monies or subsidies to specific topical areas in both science and technology output avenues.

Our method can be used in private R&D settings, with firms having a greater understanding of what potential avenues of research their in-house competences allow for, versus out-sourcing specific research-intensive tasks instead. Venture capital firms would find this method useful when determining the suitability and sustainability of the knowledge and knowledge producers involved in potential ventures. In short, this method enables a fine-grained approach to determining the applicability of past research to a future technology, which was previously not possible.

The implications of the research conducted in this thesis focus on (i) the methodologies for identifying and tracking knowledge transfer, and (ii) conditions for knowledge transfer in general and, more specifically, the Science Park as both a driver and an environment for knowledge transfer. The analysis of Chapter 5 leads to several implications for Science Parks. Science Parks represent a massive investment in financial terms, which of course will factor into technology development and knowledge and technology transfer for many, if not most, academic spin-offs. In scientific literature and policy discussions, the subject of Science Parks has typically being marked by conflicting opinions and findings. Studies routinely either conclude that Science Parks stimulate regional development or, alternatively, show no evidence of contributing to the local economy or innovative capacities of firms in them. These conflicting results stem from a lack of a consistent framework with which to evaluate Science Parks. Science Parks, in theory, should be enormous drivers of development and innovation. They provide the benefits of agglomeration, they provide a contact space for firms to interact, and they provide close proximity to a university and an educated workforce. But all too often, reasons cited by subjects in this study (and many others) for locating in Science Parks only relate to agglomeration and practical effects. Science Parks have a role in ensuring that there are enough ‘different but similar’ tenants to not only encourage the growth of firms in the park, but also the Science Park itself, and the many industrial partners and universities and government looking to see their investments being capitalised on. Each Science Park can, and should, provide an ecology of firms and knowledge that is autocatalytic in nature, providing inputs and outputs that can be utilised by most, if not all, who are located there.

6.6 References

Agrawal, A. et al. (2006). Gone But Not Forgotten: Labor Flows, Knowledge Spillovers, and Enduring Social Capital. *Journal of Economic Geography, 6*, 20.

Baba, Y. et al. (2009). How do collaborations with universities affect firms' innovative performance? The role of "Pasteur scientists" in the advanced materials field. *Research Policy, 38*(5), 756-764.

Bates, T. et al. (2004). Authorship criteria and disclosure of contributions: comparison of 3 general medical journals with different author contribution forms. *Jama, 292*(1), 86.

Carayannis, E.G. et al. (1998). High-technology spin-offs from government R&D laboratories and research universities. *Technovation, 18*(1), 1-11.

Cohen, W.M. et al., (2002). R&D spillovers, patents and the incentives to innovate in Japan and the United States. *Research Policy 31*, 1349-1367.

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

Das, T.K. & Teng, B.S. (1997). Time and Entrepreneurial Risk Behavior. *Entrepreneurship Theory and Practice, 22*(2), 69-71.

Dubine, P. & Aldrich, H. (1991). Personal and Extended Networks are Central to the Entrepreneurial Process. *Journal of Business Venturing, 6*(5), 305-313.

Elfring, T. & Hulsink, W. (2003). Networks in entrepreneurship: The case of high-technology firms. *Small business economics, 21*(4), 409-422.

Furukawa, R. & Goto, A. (2006). Core scientists and innovation in Japanese electronics companies. *Scientometrics, 68*(2), 227-240.

Greve, A. & Salaff, J.W. (2001). The development of corporate social capital in complex innovation processes. In Gabbay, S.M. & Leenders, R. (Eds.), *Social Capital of Organizations* (Vol. 18, pp. 107-134): Emerald Group Publishing Limited.

Healey, P. et al. (1986). An experiment in science mapping for research planning. *Research Policy, 15*(5), 233-251.

Karki, M. (1997). Patent citation analysis: A policy analysis tool. *World Patent Information, 19*(4), 269-272.

Koh, F.C.C. et al. (2005). An analytical framework for Science Parks and technology districts with an application to Singapore. *Journal of Business Venturing, 20*(2), 217-239.

Leydesdorff, L. et al. (1994). Tracking areas of strategic importance using scientometric journal mappings. *Research Policy, 23*(2), 217-229.

Lin, N. (1999). Building a network theory of social capital. *Connections, 22*(1), 28-51.

Löfsten, H. & Lindelöf, P. (2005). R&D networks and product innovation patterns--academic and non-academic new technology-based firms on Science Parks. *Technovation, 25*(9), 1025-1037.

Moed, H.F. (2000). Bibliometric indicators reflect publication and management strategies. *Scientometrics, 47*(2), 323-346.

Nalebuff, B.J. & Brandenburger, A.M. (1996). *Co-opetition*: HarperCollinsBusiness.

Nelson, R. (2004). The market economy, and the scientific commons. *Research Policy, 33*(3), 455-471.

Quintas, P. et al. (1992). Academic-industry links and innovation: questioning the Science Park model. *Technovation, 12*(3), 161-175.

Siegel, D.S. et al. (2003). Assessing the impact of university Science Parks on research productivity: exploratory firm-level evidence from the United Kingdom. *International Journal of Industrial Organization, 21*(9), 1357-1369.

Steffensen, M. et al. (2000). Spin-offs from research centers at a research university. *Journal of Business Venturing, 15*(1), 93-111.

Tamada, S. et al. (2006). Significant difference of dependence upon scientific knowledge among different technologies. *Scientometrics, 68*(2), 289-302.

Tang, L. & Walsh, J.P. (2010). Bibliometric fingerprints: name disambiguation based on approximate structure equivalence of cognitive maps. *Scientometrics, 84*(3), 763-784.

Westhead, P. & Batstone, S. (1998). Independent Technology-based Firms: The Perceived Benefits of a Science Park Location. *Urban Studies, 35*(12), 2197-2219.

Yank, V. & Rennie, D. (1999). Disclosure of researcher contributions: a study of original research articles in The Lancet. *Annals of internal medicine, 130*(8), 661.

Zahra, S.A. & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. *Academy of Management Review, 27*(2) 185-203.

Zucker, L.G. & Darby, M.R. (1996). Star scientists and institutional transformation: Patterns of invention and innovation in the formation of the biotechnology industry. *Proceedings of the National Academy of Sciences of the United States of America, 93*(23), 12709.

Summary

We understand technologies as the result of knowledge accumulated over time and applied in varied, and sometimes new, forms. Education and practice allow scientists and researchers to understand the phenomena they observe at a fundamental level, and to devise novel methods to apply their understanding of them.

However, the knowledge that is generated in one locale frequently needs to be translated, transferred or transliterated to find meaningful application in another. In other words, in the dynamics of science and technology, the spawning grounds of theory and the hatching grounds of application are divided by an ocean of experience and time - and it is across this ocean we aim to swim. The transfer of knowledge across the metaphorical ocean of experience and time is not radically different from the reality. The end-results of the vast interplay between individuals, firms, universities and environments - be they products, processes or ideas - follow convoluted paths. It takes a concerted effort to follow and trace these paths, be it at the fine-grained level of two individuals communicating, or at the supra-national policy level. There remains uncertainty in the research that has been produced on knowledge transfer in that many questions remain regarding the operationalisation of knowledge transfer. We still do not know what knowledge is transferred, from where and to whom, how exactly the transfer and reception work, and the conditions surrounding the transfer. In addition, this line of questioning is not only of scholarly interest, but also of interest to society in terms of innovation and innovation policy, higher education and science policy. Industry has a vested interest in this, as knowledge transfer between academia and industry provides a significant portion of the inspiration and knowledge they require to produce and develop products and services.

To deduce the processes and mechanisms involved in knowledge transfer, it is necessary to define the three primary aspects of knowledge transfer. The first involves the knowledge itself – how was it generated, how has it developed and how is it primed for transfer. The second involves the 'sender' and 'receiver' of the information or knowledge – who are they and how has each contributed to the knowledge. And the third involves the environment – how have the conditions surrounding the knowledge facilitated a productive transfer. These aspects form the basis of my primary question wherein: ***What knowledge elements are transferred from academia to industry, how are they transferred, and what factors influence this transfer?***

In researching the precise knowledge elements being transferred, the scientific publications of the person(s) under study must be positively identified as belonging to that individual and not another researcher of the same name. With the rise of the Asian science systems, and the associated low variance in Asian researcher names, this problem is likely to get worse. To tackle this, we strongly require an understanding of the problems related to name ambiguity, plus a reliable and effective process to accurately disambiguate the sometimes vast number of publications.

1 Disambiguation

Automated approaches to disambiguation are necessary and tends to follow either a computer science or a sociological/linguistic approach or a combination of the two. These approaches have

been successful to a degree but most suffer from a common drawback, that of data discarding. For example, studies utilising key words suffer if any records are missing their keywords. Another example is that of using co-author similarity to determine if two records are from the same author. When using co-authors, how do we handle records with only one author i.e. no co-authors? In practice, these records are discarded, to the detriment of the resulting precision and recall of the algorithm. To address the issue of data accuracy, the second sub-question of this thesis is: ***How can we disambiguate researchers with an effective balance between precision and recall?***

Three problems were found with current techniques: discarding of data, limited selection of metadata, and a lack of consideration for evolving research streams and actual contributions to research by authors. The solution to handling the first two problems, data discarding and limited metadata, was to select the best possible alternate combination of available metadata. In other words, for records missing one or more fields, the proportional discriminating powers of the available metadata were adjusted depending on available combinations. All this occurs 'on the fly' within the algorithm, based on previous calibrations. The last problems, that of evolution of terminologies, research streams, or knowledge homogeneity, and author contributions rely on recognising the changing roles, topics and requirements of researchers as they progress through their academic life cycle. For department heads, author status is occasionally 'honorary', in that much of the research and write-up was actually performed by others. In evaluations, however, every author listing counts. In other current techniques, similarity calculations assume input from *all* authors, in *equal* parts. Relying on this assumption leads to networks of publications by authors, exhibiting low or no similarity. This problem is tackled by assuming that contributions by authors vary according to listing (an exception is made for alphabetically ordered listings) and the discriminating powers of the indicators were adjusted to allow for this. For example, the importance of the title or abstract words (most likely selected by the authors writing the paper) is reduced, whereas the journal name (an aspect more likely to be decided on strategic grounds by the head of department) is granted more discriminatory power in the algorithm.

A similar practice is deployed to adjust for the evolution of research streams. Over time, a researcher's usage of words will change to reflect changing fields of interest and publishing. When comparing records published ten years apart, by apportioning less discriminatory power to title or abstract words but more power to journal field characterisation or cited references, one can account for the evolution of researchers' interests. Taking the dynamic selection of alternate metadata together with the adjustments for evolving research interests and contributions means that the developed algorithm allows for more accurate and inclusive results.

2 Knowledge transfer

On a practical level, knowledge transfer and associated mechanisms typically focus on mediums, examples of which include technology or skills where participants receive the knowledge required to perform tasks with a certain technology through the construction and utilisation of that technology itself. Transfer mediums are typically codified in publications and patents or can be tacit. Commonly used indicators of knowledge transfer are based on patent and publication data. Knowledge transfer has typically been addressed in the extant literature as something that occurs as matter of fact. However, there are more complex processes at work within knowledge transfer, other than merely assuming or expecting occurrence. To start, the actual knowledge elements

transferred serve as a black box and what is missing is an adequate methodology for *quantifying* the tracks and knowledge being transferred. To aid in answering the primary research question, a second sub questions is necessary, specifically ***how can we identify knowledge elements and their attributes in an operational way, and what elements are transferred between actors?***

To answer this question, a novel method was developed to analyse specific contributions by individuals and institutions, to the developmental routes of specific quanta of knowledge. This method was applied to an individual firm founder, looking at (a) how the scientific background of the patent corpus links to the scientific output of the inventor; (b) how a researcher operates in a collaborative environment, and if the contributions of those contributors are visible in the patent corpus; and (c) whether the inventor demonstrates a level of adaptive knowledge acquisition, necessary for the development of a technology.

The non-patent literature references (NPLR) found in patent applications were used to link patent and publication data. The NPLRs and the author/inventor's corpora of publications were grouped together, linking their shared title words and cited references, and then clustered. The topical foreground and cognitive background of the research cited by the patent applications, and that of the author/inventor's whole corpus of publications, provided a clear view of the required knowledge platforms and absorptive capacity of the inventor and of the contribution it makes to the development and transfer of knowledge

For further clarification, each of the inventor/author's research streams was differentiated further by the introduction of 'concept clusters'. With these concept clusters, the specific, rather than general, contributions made by the researcher involved were identified. This approach took into account the role of co-authors and co-inventors, identifying the specific expertise these collaborators added to the technologies. The institutional affiliation of the inventors and authors also gave an institutional approach to the contributions to the technologies. Additionally, the multidirectional aspect of knowledge and skill transfer between basic and applied research can be examined in minute detail using this method, including the absorptive capacity of individuals and institutions.

The primary results of this show that during the initial stages of a technology's development, the individual in the case study recognised the importance of exogenous knowledge sources. The expertise of his network of co-inventors and co-authors increased his scope and ability to source new knowledge required for the technologies. The technologies were initially based upon the individual's knowledge and skills set, but some aspects were outside his expertise. To combat this perceived gap in knowledge, the assimilation of new knowledge was necessary, and this was ultimately visible in the patent and publication analysis output. Research cited by patent applications that did not fall within the individual's or his co-inventors' expertise was quickly incorporated into the research agenda. The result of this research was seen in the increased publication output in these sub-fields and in some cases the eventual citing of this supplementary research by the patent applications in subsequent developmental stages of the technologies.

3 Absorptive capacity and academic spin-offs

Absorptive capacity may be considered both in terms of the individuals comprising the firm, and

as the firm itself. As stated by Cohen and Levinthal, "Beyond diverse knowledge structures, the sort of knowledge that individuals should possess to enhance organizational absorptive capacity is also important. Critical knowledge does not simply include substantive, technical knowledge; it also includes awareness of where useful complementary expertise resides within and outside the organization". In this manner a key aspect is the communication between the firm and the outside world. The concept of absorptive capacity was expanded on to include potential and realised absorptive capacity. These include, for potential absorptive capacity, *acquisition* – which necessitates the taking of stock or inventory of the current assets and knowledge platforms; and *assimilation* - which requires the knowledge intended to be brought in not only to be understood theoretically but also in terms of its place within current knowledge platforms. In realised absorptive capacity, the dimensions of *transformation* – which includes the ability to create novel knowledge by adding external knowledge to the current platform, and *exploitation* – in which results of the combination are brought to light. These could include, but are not limited to, patent applications, scientific publications or new work processes.

To address some of the complex processes in measuring knowledge transfer and absorptive capacity, studies frequently involve academic spin-offs because they provide the clearest identifiable path of knowledge transfer, where an idea can be followed from its inception to its commercial roll-out through a specific individual or group. Spin-offs embody an idea which was developed in academia and deemed to be commercially viable, but they require a dedicated entity to manifest. Overall, studies on spin-offs provide indications of the roles of the individuals involved with the knowledge transfer, as well as the source and end-user environments of the knowledge, but do not examine the effects of the individuals and their environments on the actual knowledge elements being transferred.

Returning to the transfer of knowledge elements, in order to link absorptive capacity and spin-offs, we examine a common route to enabling the infrastructure for absorptive capacity. For spin-offs, the environment is crucial for absorptive capacity to occur. The environment offers firms a *choice* of knowledge, and *access* to an environment is often the first step for firms stepping outside the university. For academic spin-offs, an environment that provides this is often a Science Park.

4 Science Parks

Science Parks provide an environment to promote knowledge transfer and interactions between firms, universities and small labs. They provide a contact space between the 'fast applied science' of industry and the 'slow basic science' of the university and provide a technological platform for economic development at a regional or national level.

Science park locations primarily appeal to firms which are either industry-based spin-outs, or academic spin-offs. There are three distinct reasons at the heart of the motivations of each type of firm to join a Science Park , the first of which is related to neoclassical theory in which transport, labour costs, distance to customers, and agglomeration economies are influential. The second set of reasons stem from behavioural aspects including the presence of mediators, gatekeepers or information channels in the form of the Science Park management. Additionally, the reputational advantages of situating in a Science Park play a large role in influencing firm

founders to locate in a park. Most importantly for this thesis, the third set of reasons relate to structuralist approaches, including access to an innovative, networked environment, in which the presence of a Higher Education Institution plays a central role. From this, the third and last sub-question: ***What resources, and from which actors and operational spheres, contribute most significantly to the development of an academic spin-off and its host technology?***

For firms choosing to locate within a Science Park, neoclassical location theory tends to dominate the decision processes of the firm founder. This typically includes logistical issues, such as the proximity to the founder's home. This is not to say that firms interviewed considered only these issues, but rather that *practicalities* won over *potentialities.* In interviews with firm founders, all expressed interest in being able to collaborate with other firms in the Science Park (in line with one of the espoused benefits of locating within a Science Park). However, there was little evidence in the publication and patent data to show that they actually conducted collaborative research with other firms located at the Park.

That is not to say that there was no collaboration, rather there was no substantial evidence of collaboration. From the patent and publication data, the regional and international characteristics of co-assignees and co-authors show that for almost all of the firms, collaborative activities were common, but with firms *outside* the Science Park. Academic collaborations were primarily with the local HEI, Leiden University in this case, and a few founders maintained strong links with their alma maters beyond Leiden. This was reflected in the interview data where the interactions were internal, i.e. initiated before firm's formation and location to the Science Park, and external, i.e. with academic and industrial partners elsewhere in the country and abroad.

Social capital as a resource can be considered supplementary and enabling to the stock knowledge, financial capital and skills of an entrepreneur. For the firm founders who exploited their networks, they drew capital from either internal sources (i.e. historically through personal relationships prior to firm formation) or from sources external to the Science Park. Only a few firms reported any interactions of any nature with the Science Park administration or other firms located at the Science Park. We found that the scientific capabilities of the firm founder were significant in developing and expanding the firm's scientific base, and for its eventual patent output. The substantial similarities between the patent content and the scholarly output of the firm founder (our proxy being the co-location of NPLRs and the founder's publication corpora) and the number of active research streams at and after incorporation both suggest that the scientific base of the founder had a large supporting role.

The lifeblood of a Science Park is its tenants. Science Parks compete for tenants, and those tenants exist in a *coopetitive* environment. Tenants compete for access to networks and the benefits these networks bring. They also are presented with the opportunity to cooperate with other tenants in the park, sharing resources and mitigating risks whilst expanding their potential access to networks. To counteract the diversity issues in a Science Park in terms of evaluating their utility, more emphasis is needed on analysing the knowledge structures and competences of the firms.

5 All together

The overarching research question of this thesis (*what knowledge elements are transferred from academia to industry, how are they transferred, and what factors influence this transfer?)* is at first glance a broad question. It is necessarily broad so as to encompass the complexity of knowledge transfer in relation to absorptive capacity, social capital and the environment in which these knowledge transfer processes take place. In anticipation of the inevitable question of applicability, this thesis goes some way to providing a toolbox for parties that are interested in discovering which elements are transferred, where to and where from, and what factors influence this transfer, for *their* specific field of application.

The underpinnings of this methodological toolbox begin with considering the effect of an individual's previous research on future research plans, and the similarities between past and present current research streams. The methods developed and insights provided in the chapter on disambiguation allow us to analyse factors such as the similarities and differences between research conducted during the PhD phase of a researcher's career and the professorial phase. Over time, an individual's research contributions may change with rank and eventual specialisation, but their incorporated knowledge and skill sets developed remain. Research conducted in academia and eventually applied in industry follows a convoluted path. We needed to gain an understanding of this path for our disambiguation algorithms to succeed, and such an understanding provides a first glimpse at what knowledge elements are transferred over time.

The methodology and case study in the third and fourth chapters serve as a vehicle to examine the specific contributions of an existing knowledge base to the development of a technology platform. In other words, identifying the knowledge elements and, to a certain extent, how they are transferred. The knowledge base does not necessarily come from one individual, but also from co-authors and co-inventors, and from other researchers working in different research settings. The methodology outlined in these chapters provided a toolkit for us to uncover the linkages between research conducted within academia and the eventual application of that research in industry, and the case study provided an example of what our approach can reveal. By applying our new method to a real case study, we demonstrated the ability to combine exogenously generated knowledge with a current knowledge base. New research conducted by Nakamura was guided by previous efforts, indicating that research practices and results are constantly evolving to inform, guide and provide the basis for extensions to different technologies. With detailed descriptions of the linked chains of research, we showed that a research corpus of an individual and their co-inventors and co-authors, can be readily recognised and identified in the exploitation of their research (i.e. in patents). The thematic links between the technologies and the underlying science was clearly identified in this chapter, verifying our methodological approach.

There is a long list of purported benefits for a firm to decide to locate to a Science Park. Its close proximity to a HEI and the prevalence of like-minded firms in the vicinity are examples of the reasons firms decide to locate to a Science Park. In social capital terms, a Science Park provides accessibility to resources, but for the firms in the case study, for the most part, these resources existed *in potentia*. It is important to note that for the firms under study there were no access barriers imposed by the Science Park. All the firm founders considered the Science Park to be an

important *potential* source of collaborations and customers. If the opportunity arose, all the firm founders stated they would consider it. However, using the lens of only the purported benefits of locating to a Science Park, we failed to see more than practical benefits.

It is the firms that drive the success of a Science Park, and each firm is driven by its own scientific capacities and potential market linkages. For firms to truly enjoy the network benefits of a Science Park, there should be an overlap of not only their fundamental or applied sciences, but also of potential collaborators and customers. As such, there needs to be a greater emphasis on the underlying sciences and technologies hosted by each firm when setting up a Science Park.

An important step to answering the primary research question was to consider two perspectives: access to resources and technology development. It was necessary to blend the two perspectives, as many of the opportunities in one arise from the other. Ideas generated by a scientist in academia are not initially beholden to entrepreneurial dynamics within their network. They are, however, subject to the incentives structure of academia. New research streams are common in academia, where networks of scientists contribute to one another's research, iteratively guiding the development of an idea or technology. If a certain stream or idea is deemed suitable for exploitation, the scientist/entrepreneur/founder's various networks simultaneously come into play, and the opportunities for further scientific development diminish. In their place, commercialisation of the idea becomes paramount.

Incentive structures shape both the strategies of academic researchers and industrial researchers in terms of valuing their research and results, and thus what aspects of the research will be developed and transferred. For academic spin-offs, access to scientific and technical networks and/or resources is not restricted to academia but extend into industry. The development of these resources is crucial to the development of the technology, but is secondary for the spin-off itself. It is access to the organisational, financial and regulatory networks that tops the priority list and begins to refine the technology. The characteristics of an idea conceived in academia and transferred to industry vary in many ways but, using the developed tools, we can identify its origins, track its evolutionary path, and examine the effects of the environment.

Nederlandse samenvatting

De intellectuele kringloop: overdracht en dynamiek van kennis tussen de academische wereld en bedrijfsleven

We zien technologieën als het resultaat van kennis die in de loop van de tijd is vergaard en op uiteenlopende, soms nieuwe, manieren wordt toegepast. Door opleiding en praktijkervaring verwerven wetenschappers en onderzoekers op fundamenteel niveau inzicht in de fenomenen die ze bestuderen, en kunnen ze nieuwe methoden ontwerpen om deze inzichten toe te passen.

Om de kennis die op de ene locatie wordt gegenereerd op een andere locatie op een zinvolle manier te kunnen toepassen is in veel gevallen echter een vorm van vertaling, overdracht of transliteratie nodig. Met andere woorden: in de dynamiek van wetenschap en technologie worden het continent waar de theorie ontspruit en het continent waar de toepassing tot bloei komt vaak van elkaar gescheiden door een oceaan van ervaring en tijd. En die oceaan willen we graag oversteken. De overdracht van kennis naar de andere kant van deze metaforische oceaan van ervaring en tijd is niet wezenlijk anders dan wat er in de realiteit gaande is. De eindresultaten van de complexe interactie tussen individuen, bedrijven, universiteiten en omgevingen – zowel producten, processen als ideeën – volgen kronkelige wegen. Om deze wegen te volgen en in kaart te brengen, zowel op het microniveau van twee met elkaar communicerende individuen als op het niveau van het landenoverstijgend beleid, is een gerichte inspanning nodig. Veel vragen op het gebied van de operationalisering van kennisoverdracht zijn nog onbeantwoord. Dat leidt tot onduidelijkheid in dit onderzoeksgebied. We weten nog steeds niet welke kennis wordt overgedragen, waarvandaan en aan wie, hoe de overdracht en ontvangst precies in hun werk gaan en wat de omstandigheden zijn waaronder overdracht plaatsvindt. Bovendien zijn deze vragen niet alleen van belang voor de wetenschap, maar ook voor de samenleving, in het kader van innovatie en innovatiebeleid, hoger onderwijs en wetenschapsbeleid. Het bedrijfsleven heeft hier groot belang bij omdat de kennisoverdracht tussen de academische wereld en het bedrijfsleven een zeer belangrijke rol speelt in de inspiratie en de kennis die het nodig heeft om producten en diensten te produceren en ontwikkelen.

Om te onderzoeken welke processen en mechanismen een rol spelen bij de overdracht van kennis, moeten we eerst de drie belangrijkste aspecten van kennisoverdracht definiëren. Het eerste aspect is van betrekking op de kennis zelf: hoe is deze gegenereerd, hoe is ze ontwikkeld en hoe is de overdracht ervan voorbereid? Het tweede aspect betreft de 'zender' en de 'ontvanger' van de informatie of kennis: wie zijn zij en op welke manier heeft elk van hen bijgedragen aan de kennis? Het derde aspect heeft te maken met de omgeving: op welke manier hebben de omstandigheden bijgedragen aan een productieve overdracht van de kennis? Deze aspecten

vormen de basis van mijn hoofdvraag, die luidt: ***Welke kenniselementen worden er overgedragen van de academische wereld naar het bedrijfsleven, hoe worden ze overgedragen en welke factoren zijn van invloed op deze overdracht?***

Om te kunnen onderzoeken welke kenniselementen precies worden overgedragen, moet eerst onomstotelijk worden vastgesteld dat de wetenschappelijke publicaties van de persoon (of personen) die wordt (worden) bestudeerd inderdaad aan die persoon kunnen worden toegeschreven en niet aan een andere onderzoeker met dezelfde naam. Nu Aziatische wetenschappers een steeds belangrijkere rol gaan spelen zal dit probleem, vanwege de relatief kleine variatie aan namen onder Aziatische wetenschappers, vermoedelijk alleen nog maar groter worden. Om dit aan te pakken hebben we dringend inzicht nodig in de problemen die gerelateerd zijn aan ambiguïteit op het gebied van namen en hebben we een betrouwbaar en effectief proces nodig om het soms enorme aantal publicaties van deze ambiguïteit te ontdoen.

1 Disambiguatie

Geautomatiseerde oplossingen voor disambiguatie zijn noodzakelijk. Gewoonlijk zijn deze gestoeld op een van de computerwetenschappen of een sociologisch-linguïstische benadering, of op een combinatie daarvan. Deze methoden hebben tot op zekere hoogte succes, maar kennen in de meeste gevallen een belangrijk nadeel, te weten dataverlies. Indien bijvoorbeeld met zoektermen wordt gewerkt, komen documenten waarin deze zoektermen niet voorkomen niet boven water. Een ander voorbeeld is het gebruik van overeenkomsten tussen de co-auteurs om te bepalen of twee documenten van dezelfde auteur afkomstig zijn. Hoe gaan we dan om met documenten waar slechts de naam van één auteur onder staat, met andere woorden, waarbij geen sprake is van een co-auteur? In de praktijk worden deze documenten dan buiten beschouwing gelaten, met nadelige gevolgen voor de resulterende precisie (voorspellende waarde) en recall (gevoeligheid) van het algoritme. Met het oog op dit probleem van de nauwkeurigheid van de data, luidt de eerste deelvraag van dit proefschrift: ***Hoe kunnen we onderzoekers goed van elkaar onderscheiden, met de juiste balans tussen precisie en recall?***

Geconstateerd werd dat er aan de huidige technieken drie problemen kleven: het verloren gaan van data, de beperkte selectie van metadata en het onvoldoende in beeld komen van onderzoeksstromen die nog in ontwikkeling zijn en van recente onderzoeksbijdragen van auteurs. De oplossing voor de eerste twee problemen – dataverlies en beperkte metadata - bestond uit het kiezen van de best mogelijke alternatieve combinatie van beschikbare metadata. Met andere woorden: voor documenten waarin een of meer velden ontbraken, werd het relatieve onderscheidend vermogen van de beschikbare metadata aangepast op grond van beschikbare combinaties. Dit alles gebeurt 'gaandeweg' binnen het algoritme, op grond van eerdere kalibraties. De laatste problemen, die van terminologieën en onderzoeksstromen die nog in ontwikkeling zijn, van homogeniteit van kennis, en van de bijdragen van auteurs, hangen samen met de onderkenning van de veranderende rollen, onderwerpen en eisen waar onderzoekers in de loop van hun wetenschappelijke carrière mee te maken hebben. Het auteurschap van faculteitshoofden is soms een vorm van eerbetoon, in die zin dat het onderzoek grotendeels wordt uitgevoerd door anderen, die tevens het artikel hebben geschreven. Bij evaluaties worden echter alle auteursvermeldingen meegeteld. Bij andere vaak gebruikte technieken wordt er in de similariteits-

berekeningen van uitgegaan dat alle auteurs in gelijke mate hebben bijgedragen. Vanwege deze aanname ontstaan er netwerken van publicaties door auteurs waarin niet of nauwelijks sprake is van similariteit. Dit probleem wordt aangepakt door aan te nemen dat de verschillende mate waarin auteurs hebben bijgedragen afgeleid kan worden uit de volgorde waarin hun namen worden vermeld (tenzij deze in alfabetische volgorde zijn geordend). Het onderscheidend vermogen van de indicatoren werd daarop aangepast. Zo werd bijvoorbeeld het gewicht van de titel en van de woorden van de samenvatting (over het algemeen gekozen door de auteurs die het paper geschreven hebben) verlaagd, terwijl de naam van het tijdschrift (een aspect waarover in de meeste gevallen het faculteitshoofd besluit op strategische gronden) een groter onderscheidend vermogen in het algoritme kreeg toebedeeld.

Een vergelijkbare methode werd toegepast om rekening te kunnen houden met de ontwikkeling van onderzoeksstromen. Het woordgebruik van een onderzoeker zal in de loop van de tijd veranderen tengevolge van veranderende interesse- en publicatiegebieden. Bij de vergelijking van documenten die met een tussenpoos van tien jaar zijn gepubliceerd, kan recht worden gedaan aan de ontwikkeling van de interessegebieden van de onderzoeker door minder onderscheidend vermogen toe te kennen aan de titel of de woorden van de samenvatting, en meer aan de referenties of het vakgebied waarop het tijdschrift betrekking heeft. Door de dynamische selectie van alternatieve metadata, in combinatie met aanpassingen voor veranderende onderzoeksinteresses en -bijdragen, is een algoritme ontwikkeld dat meer nauwkeurige en completere resultaten oplevert.

2 Kennisoverdracht

In de praktijk is de aandacht bij kennisoverdracht en de daaraan gerelateerde mechanismen doorgaans gericht op de middelen waarmee deze overdracht plaatsvindt, zoals technologie of vaardigheden, waarbij deelnemers de kennis ontvangen die nodig is om bepaalde taken met een bepaalde technologie uit te kunnen voeren door middel van constructie en gebruik van die technologie zelf. Deze middelen van kennisoverdracht zijn in publicaties en octrooien veelal gecodificeerd of blijven impliciet. De meest gebruikte indicatoren van kennisoverdracht zijn gebaseerd op gegevens uit patenten en publicaties.

Kennisoverdracht wordt in de huidige literatuur over het algemeen benaderd als iets dat voor zich spreekt. Binnen kennisoverdracht zijn echter complexere mechanismen in het spel, naast de simpele aanname of verwachting dat het gebeurt. Ten eerste functioneren de kenniselementen die worden overgedragen als een black box en ontbreekt een adequate methodologie voor het *kwantificeren* van de kennis die wordt overgedragen en de wegen waarlangs dat gebeurt. Om de hoofdvraag te kunnen beantwoorden, is een tweede deelvraag nodig: ***hoe kunnen we het achterhalen van kenniselementen en de eigenschappen ervan operationaliseren en welke elementen worden er tussen de actoren overgedragen?***

Om deze vraag te beantwoorden is een nieuwe methode ontwikkeld waarmee specifieke bijdragen van individuen en instellingen aan de ontwikkelingsroutes van specifieke kennisclusters kunnen worden onderzocht. Deze methode werd toegepast op een individuele oprichter van een bedrijf. Daarbij werd gekeken (a) hoe het verband is tussen de wetenschappelijke achtergrond van het patentcorpus en de wetenschappelijke output van de uitvinder; (b) hoe een onderzoeker binnen een samenwerkingsomgeving opereert en of de bijdragen van iedereen zichtbaar zijn in

het patentcorpus; en (c) of de uitvinder in enige mate blijk geeft van een adaptieve manier van kennisvergaring, die vereist is voor de ontwikkeling van een technologie.
De verwijzingen in niet-patent literatuur (non-patent literature references, NPLR's) uit de patentaanvragen werden gebruikt om patent- en publicatiegegevens aan elkaar te koppelen. De NPLR's en de publicatiecorpora van de auteur/uitvinder werden in samenhang met elkaar gegroepeerd op grond van overeenkomsten tussen titelwoorden en verwijzingen, en vervolgens geclusterd. Op basis van het onderwerp en de cognitieve achtergrond van het in patentaanvragen aangehaalde onderzoek en die van het hele publicatiecorpus van de auteur/uitvinder, ontstond een duidelijk inzicht in de voor de uitvinder vereiste kennisplatforms en het vereiste absorptie-vermogen, en in de rol die deze spelen bij de ontwikkeling en overdracht van kennis.

Om nog meer helderheid te verschaffen, werd elk van de onderzoeksstromen van de uitvinder/auteur verder gedifferentieerd door 'conceptclusters' te introduceren. Met behulp van deze conceptclusters konden de specifieke (en niet zozeer de algemene) bijdragen van elke onderzoeker worden aangewezen. Door deze aanpak kon in kaart worden gebracht wat de rol van de co-auteurs en co-uitvinders was, en welke specifieke expertise elk van deze medewerkers had toegevoegd aan de technologieën. Omdat de uitvinders en auteurs aan een instelling waren verbonden, kregen de bijdragen aan de technologieën ook een instellingscomponent.
Bovendien kan met behulp van deze methode het multidirectionele karakter van de overdracht van kennis en vaardigheden tussen fundamenteel en toegepast onderzoek zeer gedetailleerd worden onderzocht, inclusief het absorptievermogen van individuen en instellingen.

De belangrijkste resultaten laten zien dat het individu uit de casestudie in het beginstadium van de ontwikkeling van een technologie het belang onderkende van exogene kennisbronnen. De invloedssfeer van deze persoon en zijn vermogen om nieuwe, voor de technologieën benodigde, kennis aan te boren, werden vergroot door de expertise van zijn netwerk van co-uitvinders en co-auteurs. De technologieën waren in eerste instantie gebaseerd op het geheel van kennis en vaardigheden van dat individu, maar sommige aspecten vielen buiten zijn expertise. Om de geconstateerde lacune in zijn kennis op te lossen, moest hij nieuwe kennis assimileren en dit werd later zichtbaar in de resultaten van de patent- en publicatieanalyse. Onderzoek waarnaar in patentaanvragen werd verwezen en dat niet tot de expertise van het individu of zijn co-uitvinders behoorde, werd snel opgenomen op de onderzoeksagenda. Het resultaat van dit onderzoek was zichtbaar in de toegenomen publicatieoutput op deze subgebieden, en in bepaalde gevallen in het feit dat in latere ontwikkelingsstadia van de betreffende technologie in patentaanvragen werd verwezen naar dit aanvullend onderzoek.

3 Absorptievermogen en academische spin-offs

Absorptievermogen kan zowel van betrekking zijn op de individuen die bij een bedrijf werken als op het bedrijf zelf. In de woorden van Cohen en Levinthal: "Behalve de verschillende kennisstructuren is ook het soort kennis waarover individuen moeten beschikken van groot belang voor het versterken van het absorptievermogen van een organisatie. Kritieke kennis bestaat niet enkel uit inhoudelijke, technische kennis, maar omvat ook inzicht in de vraag waar nuttige aanvullende expertise te vinden is, zowel binnen als buiten de organisatie." In dit verband is de communicatie tussen het bedrijf en de buitenwereld van cruciaal belang. Het concept absorptievermogen werd verbreed zodat zowel potentieel als gerealiseerd absorptievermogen eronder vallen. Voor

potentieel absorptievermogen gaat het dan om *acquisitie* (waarvoor een inventaris nodig is van de bestaande kennisplatforms en van wat er in huis is) en *assimilatie* (waarvoor vereist is dat men de kennis die men binnen wil halen niet alleen theoretisch begrijpt maar ook begrijpt welke plaats deze kennis inneemt binnen de bestaande kennisplatforms). Gerealiseerd absorptievermogen omvat de dimensies *transformatie* (waaronder het vermogen om nieuwe kennis te genereren door externe kennis toe te voegen aan het bestaande platform) en *exploitatie* (dat kijkt naar de resultaten van de combinatie). Voorbeelden van exploitatie zijn onder andere, maar niet uitsluitend, patentaanvragen, wetenschappelijke publicaties en nieuwe werkprocessen.

Met het oog op de complexe processen die aan de orde zijn bij het meten van kennisoverdracht en absorptievermogen, richt onderzoek zich vaak op academische spin-offs omdat daar het pad van kennisoverdracht het duidelijkst te traceren is, door een idee te volgen vanaf zijn ontstaan tot aan de commerciële uitrol door een bepaald individu of een bepaalde groep. Spin-offs geven gestalte aan een idee dat binnen de academische wereld is ontwikkeld en levensvatbaar wordt geacht. Er is echter een gerichte inspanning nodig om een spin-off te verwezenlijken. Onderzoeken naar spin-offs geven weliswaar een indicatie van de rol die de bij de kennisoverdracht betrokken individuen spelen en van de omgeving van zowel de bron als de eindgebruikers van de kennis, maar leveren geen inzicht op in de effecten van individuen en omgevingen op de kenniselementen die daadwerkelijk worden overgedragen.

Terugkerend naar de overdracht van kenniselementen onderzoeken we, om absorptievermogen aan spin-offs te koppelen, een gebruikelijke route waarlangs de infrastructuur voor absorptievermogen wordt versterkt. Voor spin-offs is de omgeving van cruciaal belang voor het ontstaan van absorptievermogen. De omgeving biedt bedrijven keuzemogelijkheden op het gebied van kennis, en toegang tot een omgeving is vaak de eerste stap voor bedrijven die zich buiten de universiteit willen begeven. Voor academische spin-offs is zo'n omgeving in veel gevallen een Science Park.

4 Science Parks

Science Parks vormen een omgeving die bevorderend is voor de kennisoverdracht en interactie tussen bedrijven, universiteiten en kleine laboratoria. En ze bieden een ruimte waarbinnen contact kan worden gemaakt tussen de 'snelle toegepaste wetenschap' van het bedrijfsleven en de 'langzame fundamentele wetenschap' van de universiteit. Ze bieden een technologisch platform voor economische ontwikkeling op regionaal en nationaal niveau.

De locaties van Science Parks zijn met name aantrekkelijk voor bedrijven die ofwel spin-outs zijn uit het bedrijfsleven ofwel academische spin-offs. De redenen om zich in een Science Park te vestigen verschillen naar bedrijfssoort en vallen uiteen in drie hoofdredenen. Allereerst is er de motivatie die samenhangt met de neoklassieke theorie, waarin transport, arbeidskosten, afstand tot de klant en agglomeratievoordelen van invloed zijn. De tweede set redenen hangen samen met gedragskenmerken, zoals de aanwezigheid van bemiddelaars, 'poortwachters' of informatiekanalen, in de vorm van het management van het Science Park. Verder spelen bij oprichters van een bedrijf de voordelen die vestiging in een Science Park voor de reputatie van het bedrijf hebben een grote rol bij hun besluit om zich in zo'n park te vestigen. Het belangrijkste in het kader van dit proefschrift is dat deze derde set van redenen samenhangt met structuralistische benaderingen, waaronder toegang tot een innovatieve netwerkomgeving, waarin de aanwezig-

heid van een instelling voor hoger onderwijs een centrale rol speelt. Dit leidt tot de derde en laatste deelvraag: ***Welke middelen, en vanuit welke actoren en operationele gebieden, leveren de belangrijkste bijdrage aan de ontwikkeling van een academische spin-off en zijn belangrijkste technologie?***

Voor bedrijven die besluiten om zich in een Science Park te vestigen, is de neoklassieke vestigingstheorie vaak van doorslaggevende betekenis bij de besluitvormingsprocessen van de oprichter van het bedrijf. Het gaat daarbij om logistieke afwegingen, zoals de nabijheid van de woonplaats van de oprichter. Dat wil niet zeggen dat de geïnterviewde bedrijven uitsluitend deze aspecten in hun afwegingen meenamen, maar wel dat praktische overwegingen meer de overhand hadden dan potentiële kansen. Alle geïnterviewde oprichters van bedrijven gaven aan dat ze geïnteresseerd waren in samenwerking met andere bedrijven in het Science Park (een van de geaccepteerde voordelen van vestiging in een Science Park). Uit de gegevens in de publicaties en patenten blijkt echter niet of nauwelijks dat ze ook daadwerkelijk met andere bedrijven in het park samenwerkten bij onderzoek.

Dat wil niet zeggen dat er helemaal geen sprake was van samenwerking, maar wel dat er geen aanwijzingen waren voor een grote mate van samenwerking. Op grond van de gegevens van de patenten en publicaties, blijkt uit de regionale en internationale kenmerken van de aanvragers en co-auteurs dat bijna alle bedrijven veel samenwerken, maar met bedrijven buiten het Science Park. Samenwerking op academisch gebied vond hoofdzakelijk plaats met de lokale instelling voor hoger onderwijs, in dit geval de Universiteit van Leiden. Enkele oprichters onderhielden sterke banden met hun alma maters buiten Leiden. Dit kwam ook naar voren uit de interviewgegevens, waar de interacties intern waren (d.w.z. dat ze geïnitieerd werden voordat het bedrijf werd opgericht en zich op het Science Park vestigde), en extern (d.w.z. met partners uit de academische wereld en het bedrijfsleven elders in het land en in het buitenland).

Sociaal kapitaal als middel kan worden gezien als aanvullend op en bevorderend voor de beschikbare kennis, het financieel kapitaal en de vaardigheden van een ondernemer. Bedrijfsoprichters die gebruik maakten van hun netwerk, haalden kapitaal ofwel uit interne bronnen (d.w.z. historisch op grond van persoonlijke relaties die al bestonden voordat het bedrijf werd opgericht) ofwel uit bronnen van buiten het Science Park. Slechts een paar bedrijven maakten melding van interactie met het management van het Science Park of met andere bedrijven die in het Science Park gevestigd waren. Uit ons onderzoek bleek dat de wetenschappelijke kennis en vaardigheden van de oprichter van het bedrijf een zeer belangrijke rol spelen bij het ontwikkelen en uitbreiden van de wetenschappelijke basis van het bedrijf, en voor het aantal uiteindelijk aangevraagde patenten. Zowel de zeer grote overeenkomsten tussen de inhoud van de patenten en de wetenschappelijke publicaties van de oprichter van het bedrijf (op grond van ons criterium van co-locatie van NPLR's en het totale publicatiecorpus van de oprichter) als het aantal actieve onderzoeksstromen op het moment van en na oprichting wijzen erop dat de wetenschappelijke basis van de oprichter hieraan in belangrijke mate bijdraagt.

De gebruikers vormen het kloppend hart van een Science Park. Science Parks beconcurreren elkaar om gebruikers naar zich toe te halen, en die gebruikers functioneren in een *competitieve* en *coöperatieve* omgeving. Gebruikers beconcurreren elkaar om het gebruik van netwerken en de

voordelen die deze netwerken met zich meebrengen. Ook krijgen zij de kans met andere gebruikers van het park samen te werken, middelen te delen en risico's te spreiden, terwijl zij hun potentiële toegang tot netwerken uitbreiden. Om de problemen van diversiteit binnen Science Parks aan te pakken, met het oog op de beoordeling van het nut van deze parken, moet er meer aandacht komen voor de analyse van de kennisstructuren en de competenties van de bedrijven.

5 Alles overziend

De hoofdvraag van dit onderzoek (Welke kenniselementen worden overgedragen van de academische wereld naar het bedrijfsleven, hoe worden ze overgedragen en welke factoren zijn van invloed op deze overdracht?) is op het eerste gezicht een brede vraag. Het was nodig deze vraag zo breed te formuleren, om te zorgen dat deze ruimte biedt aan de complexiteit van kennisoverdracht in relatie tot absorptievermogen, sociaal kapitaal en de omgeving waarbinnen deze processen van kennisoverdracht plaatsvinden. Vooruitlopend op de onontkoombare vraag naar toepasbaarheid, geeft dit proefschrift een aanzet tot een toolbox voor partijen die voor hun *eigen* specifieke toepassingsgebeid geïnteresseerd zijn te ontdekken welke elementen worden overgedragen en waarheen en waarvandaan, en welke factoren van invloed zijn op deze overdracht.

De basis van deze methodologische toolbox wordt gevormd door een analyse van het effect van eerder door het betreffende individu uitgevoerd onderzoek op zijn of haar toekomstige onderzoeksplannen en de overeenkomsten tussen onderzoeksstromen uit verleden en heden. Met de methoden die in het hoofdstuk over disambiguatie worden ontwikkeld, alsmede de inzichten die daar aan de orde komen, kunnen we factoren analyseren zoals de verschillen en overeenkomsten tussen onderzoek dat in de promotiefase van de carrière van de onderzoeker is uitgevoerd en onderzoek dat tijdens zijn of haar professionele carrière is uitgevoerd. De onderzoeksbijdragen van een individu kunnen in de loop van de tijd veranderen door zijn of haar academische positie en uiteindelijke specialisatie, maar het geheel aan kennis en vaardigheden waarover hij of zij beschikt, blijft behouden. Onderzoek dat binnen de academische wereld wordt uitgevoerd en vervolgens wordt toegepast in het bedrijfsleven volgt een kronkelig pad. Om succesvolle algoritmes voor disambiguatie te kunnen ontwikkelen, moeten we inzicht verwerven in dit pad. Dat inzicht geeft ons een eerste indruk van welke kenniselementen er in de loop van de tijd worden overgedragen.

De methodologie en de casestudie in het derde en vierde hoofdstuk dienen als middel voor het onderzoeken van de specifieke bijdragen van een bestaande kennisbasis aan de ontwikkeling van een technologieplatform, waarbij wordt geïdentificeerd welke kenniselementen zijn betrokken en, tot op zekere hoogte, hoe deze worden overgedragen. De kennisbasis is niet per definitie afkomstig van één individu, maar ook van co-auteurs en co-uitvinders, en van andere onderzoekers die in andere onderzoekssettings werken. De methodologie die in deze hoofdstukken is beschreven, vormde voor ons een toolkit waarmee we de verbanden konden blootleggen tussen onderzoek dat binnen de academische wereld wordt uitgevoerd en de uiteindelijke toepassing van dat onderzoek in het bedrijfsleven. Aan de hand van de casestudie hebben we laten zien wat onze aanpak aan het licht kan brengen. Door onze nieuwe methode op een echte casestudie toe te passen, hebben we aangetoond dat hiermee exogeen gegenereerde kennis gecombineerd kan worden met een bestaande kennisbasis. Nieuw onderzoek dat door Nakamura werd uitgevoerd, vond plaats op grond van eerdere onderzoeksinspanningen, waaruit blijkt dat onderzoeks-

praktijken en -resultaten voortdurend in beweging zijn en invloed hebben op, richting geven aan en de basis vormen van uitbreiding naar verschillende technologieën. Aan de hand van gedetailleerde beschrijvingen van met elkaar samenhangende onderzoeksketens, hebben we laten zien dat het onderzoekscorpus van een individu en zijn of haar co-uitvinders en co-auteurs goed herkenbaar en herleidbaar is in de exploitatie van hun onderzoek (d.w.z. in patenten). De thematische samenhang tussen de technologieën en de onderliggende wetenschap werd in dit hoofdstuk duidelijk aangetoond, waarmee de validiteit van onze methodologische benadering is onderbouwd.

Er is een lange lijst van verwachte voordelen voor bedrijven om zich in een Science Park te vestigen. De nabijheid van een instelling voor hoger onderwijs en de aanwezigheid van verwante bedrijven in de buurt zijn enkele voorbeelden van de redenen waarom bedrijven besluiten zich in een Science Park te vestigen. In termen van sociaal kapitaal verschaft een Science Park toegang tot middelen, maar voor het merendeel van de bedrijven uit de casestudie bestonden deze middelen slecht in potentie. Het is van belang op te merken dat er voor de bestudeerde bedrijven geen barrières voor toegankelijkheid werden opgeworpen door het Science Park. Alle oprichters van bedrijven beschouwden het Science Park als een belangrijke *potentiële* bron van samenwerking en klanten. Alle bedrijfsoprichters verklaarden dat zij de kans in overweging zouden nemen indien deze zich weer zou voordoen. Uitsluitend door de bril van de verwachte vestigingsvoordelen, zagen we echter niet meer dan praktische voordelen.

Het succes van een Science Park wordt bepaald door de bedrijven en het succes van elk bedrijf hangt af van zijn eigen wetenschappelijke capaciteiten en relaties tot potentiële markten. Om daadwerkelijk de voordelen te genieten van het netwerk van een Science Park, moet er niet alleen een overlap zijn tussen hun fundamentele of toegepaste wetenschappen, maar ook tussen potentiële medewerkers en klanten. In verband daarmee moet er bij het opzetten van een Science Park meer aandacht komen voor de wetenschappen en technologieën waar elk bedrijf op gebaseerd is.

Voor het beantwoorden van de hoofdvraag van dit onderzoek was een belangrijke stap het in aanmerking nemen van twee perspectieven: toegang tot middelen en ontwikkeling van technologie. Het was nodig om de twee perspectieven in elkaar over te laten lopen omdat veel van de kansen die zich op het ene gebied voordeden, voortkwamen uit het andere. Ideeën die door een wetenschapper in de academische wereld worden gegenereerd, zijn in eerste instantie niet afgestemd op de ondernemersdynamiek binnen hun netwerk. Ze worden echter beïnvloed door de beloningsstructuur van de academische wereld. Nieuwe onderzoeksstromen zijn heel gewoon in de academische wereld, waar netwerken van wetenschappers bijdragen aan elkaars onderzoek, dat vervolgens richting geeft aan de ontwikkeling van een idee of technologie. Als een bepaalde stroom of idee geschikt wordt geacht voor exploitatie, komen de verschillende netwerken van de wetenschapper/ondernemer/oprichter tegelijkertijd in actie en nemen de kansen voor verdere wetenschappelijke ontwikkeling af. In plaats daarvan komt de commercialisering van het idee voorop te staan.

Beloningsstructuren zijn bepalend voor de strategieën van onderzoekers in zowel de academische wereld als het bedrijfsleven, als het gaat om de waardering van hun onderzoeken en hun

resultaten. Ze bepalen daarmee welke aspecten van het onderzoek verder worden ontwikkeld en overgedragen. Voor academische spin-offs is toegang tot wetenschappelijke en technische netwerken en/of bronnen niet beperkt tot de academische wereld, maar ook van toepassing op het bedrijfsleven. De ontwikkeling van deze bronnen is cruciaal voor de ontwikkeling van de technologie, maar is voor de spin-off zelf secundair. Bovenaan de lijst van prioriteiten staat toegang tot de organisatorische, financiële en regelgevende netwerken Deze toegang is ook bepalend voor de verfijning van de technologie. Er zijn weliswaar verschillen tussen de kenmerken van een idee zoals het in de academische wereld ontstaat en die van het idee zoals het wordt overgedragen aan het bedrijfsleven, maar met behulp van de ontwikkelde instrumenten kunnen we de oorsprong van dat idee vaststellen, het ontwikkelingspad ervan traceren en de effecten die de omgeving erop heeft onderzoeken.

A Word of Thanks

Researching, writing, editing, printing and, in due course, defending this PhD dissertation has been a truly remarkable journey. At each step, numerous people have been involved in guiding, shaping and supporting my research and accompanying activities.

I would like to thank the committee for their valuable critique of this dissertation. Your time spent on reading this work and formulating thought-provoking challenges is of immense value to both the content of the dissertation and my own intellectual development. Some of you I know personally, and would like to thank you for your always stimulating conversations at the various occasions and conferences I've had the honour to both present and attend. For the other members of the committee, I would like to think this dissertation is a first step to getting to know each other on a more personal level.

This research would not be possible without the guidance of my promoter, Peter van den Besselaar. You provided me a space and mindset to explore the convoluted world of our field and academia in general. Our relationship has been honest, fruitful and exciting, and of course not without frustration – just as any model relationship should be. As my former employer as the previous head of SciSA, you granted me the opportunity to work at the Rathenau, and introduced me to this fascinating world. As my promoter you crystallised my thoughts and theories in a way I hope you have the opportunity to do for many others. I certainly hope we'll be able to find an exciting new vein of research to explore and mine in the future – I'll even bring the ropes and torches.

Edwin Horlings has provided much of the day-to-day support at both the Rathenau and in my PhD. I would like to thank you for being a perfect sparring partner and mentor, from my beginnings at the Rathenau to the final days of preparing this dissertation. You have the ability to stimulate those around you in a manner I've rarely seen, and wish you nothing but the best in the future.

The Rathenau Instituut has been a wonderful place to work. I would like to thank the institute for allowing me to conduct this research, and for providing an environment where creativity can flourish. Jan Staman, Barend van der Meulen and Frans Brom have crafted a workplace where great ideas are born, and I am very grateful I could be a part of it.

Within the institute, I would like to thank Clara Kemper from the communications team for providing the help necessary to turn this research into a wonderful book. You are much appreciated.

To the colleagues who recently completed their dissertations - Laurens, Marije and Elizabeth to name but a few, I would like to thank you for your advice on the whole PhD process. You provided me, and other colleagues also in the final stages of their PhDs, great directions to avoid the traps and snares, making this process much easier and more enjoyable. We all thank you for that.

The JuScis of the Rathenau have been enormously helpful and supportive. Tjerk, Rosalie, Pieter, Bei, Keelie, Stefan and Pleun have a special place in the Rathenau. For those leaving the Rathenau, I wish you all the luck in your new positions. For those staying, keep that JuSci feeling alive. To Stefan and Pleun – you are possibly the best colleagues and friends imaginable. We managed to find that perfect blend between inspiration, care and critique to help in all areas of our lives. Our coffee breaks will stand out in my memory for many, many years and I look forward to many more, no matter where we are.

Outside of the Rathenau, I would like to thank Loet Leydesdorff for introducing me to scientometrics and bibliometrics. His work has inspired many, and it has truly been an honour to have worked with you, in your role as my Master's supervisor and at the various conferences and seminars around the world. I would also like to thank Antoine Schoen and Daniel Pardo for providing much of the impetus and source data for my research. It has been a pleasure to work with you.

On the personal side, I would like to thank my friends, particularly Ruarri, Jurjen, Matt and Dave for keeping my PhD from consuming me. Your wit and relentless jokes have provided a perfect counterpoint to the seriousness of the PhD, making life and work easier in many ways, and always reminding me to never take myself too seriously. There'll be no more referencing obscure papers, I promise.

My family have provided tremendous support – in more ways than one. To my mother and father, I will always be grateful – for imbuing me with an enormous appreciation for all the curiosities the world has to offer. You each taught me to appreciate the art and science of nature and design in your own way and I love you for that. This PhD would not be possible without your love and support. To Penny, Brian, Tanya and Mikaila, you are much loved and appreciated.

I owe much of my happiness in the Netherlands to the other half of my family. Rick and Will, thank you for everything you have done for me. My life here would not be possible without you. Rikkert, Anne, Elise, Kees, Mees, Fien and Sam, you make great brothers, sisters and cousins. Thank you!

Lastly, I would like to thank Joffra. You have been nothing but patient, loving and stimulating for all the time we've been together. From our time in Cape Town, from Oost to West, this journey has been incredible and it's all thanks to you. Je bent de mooiste, liefste en slimste vrouw in de hele wereld. This PhD is for you.

Who was Rathenau?

The Rathenau Instituut is named after Professor G.W. Rathenau (1911-1989), who was successively professor of experimental physics at the University of Amsterdam, director of the Philips Physics Laboratory in Eindhoven, and a member of the Scientific Advisory Council on Government Policy. He achieved national fame as chairman of the commission formed in 1978 to investigate the societal implications of micro-electronics. One of the commission's recommendations was that there should be ongoing and systematic monitoring of the societal significance of all technological advances. Rathenau's activities led to the foundation of the Netherlands Organization for Technology Assessment (NOTA) in 1986. On 2 June 1994, this organization was renamed 'the Rathenau Instituut'.

In the dynamics of science and technology, the spawning grounds of theory and the hatching grounds of application are divided by an ocean of experience and time - and it is across this metaphorical ocean we aim to swim. In reality, products, processes or ideas are the end-results of the vast interplay between individuals, firms, universities and environments. It takes concerted effort to follow and trace these paths, be it at a fine-grained level of two individuals communicating, or a supra-national policy level. In the research that has been produced on knowledge transfer, many questions remain regarding the operationalisation of knowledge transfer. We still do not know what knowledge is transferred, from where and to whom, how exactly the transfer and reception work, and the conditions surrounding the transfer. The research question of this study, 'What knowledge elements are transferred from academia to industry, how are they transferred, and what factors influence this? aims to provide a methodological toolbox to address this.

Key results of this research address the concurrent nature of knowledge transfer, specifically the data employed to measure knowledge transfer, access to resources by actors when creating and disseminating knowledge, and the environment in which knowledge transfer processes occur.

The lines of questioning and research provided in this study are of interest to industry, and this study addresses the value to society in terms of innovation and innovation policy, higher education and science policy.

ISBN 978-90-77364-50-5
Laser Proof
9 789077 364505