



Universiteit
Leiden
The Netherlands

Early-stage detection of breakthrough-class scientific research : using micro-level citation dynamics

Winnink, J.J.

Citation

Winnink, J. J. (2017, February 22). *Early-stage detection of breakthrough-class scientific research : using micro-level citation dynamics*. Retrieved from <https://hdl.handle.net/1887/46101>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/46101>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/46101> holds various files of this Leiden University dissertation.

Author: Winnink, J.J.

Title: Early-stage detection of breakthrough-class scientific research : using micro-level citation dynamics

Issue Date: 2017-02-22

Early-stage detection of breakthrough-class scientific research

Using micro-level citation dynamics

Jos Winnink

Early-stage detection of breakthrough-class scientific research using micro-level citation dynamics

© Jos Winnink, 2017

Cover design and layout: Jos Winnink

Photograph: Scheveningen beach[®], photographed by Jos Winnink, 31 December 2015

Printing: Eburon Academic Publishers, Delft

ISBN: 978-94-6301-116-7

Early-stage detection of breakthrough-class scientific research

Using micro-level citation dynamics

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op woensdag 22 februari 2017
klokke 15:00 uur

door

Jacques Johannes Winnink

geboren te Den Haag
in 1953

Promotores

Prof. dr. A.F.J. van Raan

Prof. dr. R.J.W. Tijssen

Promotiecommissie

Prof. dr. C.A. van Bochhove

Prof. dr. D.D. Breimer

Prof. dr. K. Frenken (*Universiteit Utrecht*)

Prof. dr. J.W.M. Frenken

Voor

Annemarie

Matthijs & Thomas

Max (†) & Jo

Dankwoord (acknowledgement in Dutch)

Ik had dit onderzoek, waar dit proefschrift verslag van doet, niet kunnen voltooien zonder de steun en inspiratie van velen. Om in de termen van Marten Toonder te spreken het was *mijn eigen eenzame weg* waarop zo nu en dan de Zwarte Zwadderneel verscheen om een te veel aan opgewektheid over bereikte resultaten terecht te wijzen. Iedereen die op enige wijze heeft bijgedragen aan het succesvol afronden van mijn promotieonderzoek wil ik dan ook van harte bedanken. Het is ondoenlijk om iedereen met name te noemen, maar voor enkelen wil ik een uitzondering maken.

Allereerst ben ik Merel Heimens Visser en Guus Broesterhuizen veel dank verschuldigd omdat zij binnen het Octrooicentrum Nederland (OCNL) de faciliteiten hebben geschapen waardoor ik dit promotietraject heb kunnen uitvoeren. Al snel kom ik dan bij het Centrum Regionale Kennisontwikkeling van de Universiteit Leiden, dat nu door het leven gaat als FGGA-CRK/Dual PhD Centre. Dit centrum bevond zich destijds nog in een embryonale fase. Zonder andere medewerkers en mijn collega duale promovendi tekort te willen doen, gaat mijn speciale dank uit naar de oprichter en toenmalige directeur Adriaan in 't Groen. Met veel genoegen denk ik met name terug aan de inspirerende beginperiode in het Koesthuis van Lange Voorhout 44. Op deze locatie heb ik met Adriaan vele gesprekken gevoerd over promoveren en aanverwante zaken. Ook toen ik later meer tijd doorbracht in Leiden bij het CWTS dan bij het CRK bleven deze gesprekken doorgaan zij het met een lagere frequentie. Adriaan bedankt voor deze *bildungs*-ervaring.

Het gehele promotietraject heb ik als een periode van *bildung* ervaren. Maar voorafgaand aan het promotieonderzoek was er nog een andere vorm van *bildung*. Hierin hebben mijn vrienden van de judovereniging Shi-Han in Leidschendam André Nieuwenhout, Ivo Zweekhorst en Paul de Witt een grote rol gespeeld. Maar bovenal gaat mijn dank uit naar mijn judo-leraar de heer Verhagen, die ondanks het feit dat ik pas op 39-jarige leeftijd begon met judo, er in is geslaagd om mij in een periode van 13 jaar te begeleiden van 0-niveau naar eerste Dan. Dit traject ging gepaard met het aanbrengen van ook voor een promotie onontbeerlijke eigenschappen als incasserings- en doorzettingsvermogen alsmede de instelling om het wezen –de kern– van een onderwerp, in dit geval judotechnieken, te doorgronden.

Natuurlijk gaat mijn dank ook uit naar mijn directe collega's Marcel Seip, Philip Oomen en Tom Stoop van het Cluster KOB van OCNL. Zij moesten een groter deel van de werklast dragen door mijn verminderde en ook minder regelmatige aanwezigheid.

De medewerkers van het CWTS wil ik bedanken voor het feit dat zij mij, ondanks het feit dat ik niet-fulltime aanwezig was, van meet af aan het gevoel hebben gegeven dat ik welkom was en in het bijzonder wil ik Bert van der Wurff bedanken. Met Bert heb ik een aantal jaren een kamer gedeeld en hij werd meer dan een *doorsnee* kamergenoot.

Naderend tot mijn '*inner circle*' wil ik allereerst mijn ouders bedanken die mij altijd hebben gestimuleerd en gesteund, ook op momenten dat zaken

wat minder voorspoedig verliepen. Ik kom dan uit bij mijn gezin Matthijs, Thomas en Annemarie. Altijd bleven zij geïnteresseerd in de voortgang van het onderzoek, hoewel ze zich soms afvroegen *of het onderzoek nu nog niet gedaan was*. Annemarie voor jou was het afgelegde traject misschien nog wel zwaarder dan voor mij, omdat ik (te) weinig tijd voor je had en daarnaast ook niet altijd even goed aanspreekbaar was omdat mijn prioriteiten bij het promotieonderzoek lagen. Jouw niet aflatende steun in deze periode heeft zeker bijgedragen aan de succesvolle afronding.

Abstract

The research presented in this PhD thesis describes ways of identifying at an early-stage, 2–3 years after their publication, discoveries in science that are expected to have a major impact on science. The focus in this study is on the nature of those breakthroughs, and more specifically how such phenomena can be identified at early-stage more systematically and on a large-scale. Bibliographic information extracted from those scientific publications is analysed to select patterns that may identify such ‘scientific breakthroughs’. A distinction is made between different types of breakthroughs. Given the small time-window for observations, a major methodological issue is the differentiation between discoveries that actually achieve a long-term major impact on science and those where the impact has not occurred within the first three years. These impacts are measured in terms of ‘citations’, references from subsequent research publications to the ‘target publication’ in which the discovery is documented.

This thesis begins with introducing conceptual framework that is used in the analysis of four case studies, each study focussing on a specific well-documented breakthrough discovery. The citation impact patterns are analysed in a search for bibliographic markers that are characteristic for a breakthrough discovery. The significance of the various markers, one set of markers for each type of breakthrough, were tested in a large-scale validation study.

This study resulted in several early-stage identification algorithms. Five of these algorithms were implemented as automated computerized search methods. The algorithms have proved able, retrospectively, to identify early-stage publications that present (potential) breakthrough discoveries.

Contents

1	Introduction	1
1.1	General context	1
1.2	Theoretical and conceptual framework	4
1.3	Analytical framework	13
1.4	Introduction of the studies in this thesis	17
1.5	Information sources	19
2	Case study #1: Integrase Inhibitors	25
2.1	Introduction	26
2.2	Analytical framework	26
2.3	Results	30
2.4	Discussion and concluding remarks	38
2.5	Acknowledgement	40
3	Case study #2: Graphene	41
3.1	Introduction	42
3.2	The graphene breakthrough	49
3.3	Expert opinion	59
3.4	Discussion and concluding remarks	60
3.5	Acknowledgements	62
3.A	Supplementary information — co-occurrence term maps	64
4	Case study #3: Ubiquitin	65
4.1	Introduction	66
4.2	Hypothesis, research questions, data and method	69
4.3	Results	70
4.4	Discussion	80
4.5	Concluding remarks	83
4.6	Acknowledgements	84
4.A	Supplementary material	84

5 Case study #4: Introns	89
5.1 Introduction	89
5.2 Conceptual framework	91
5.3 Methodology	92
5.4 Results	92
5.5 Concluding remarks	97
6 Validation study	99
6.1 Introduction	100
6.2 Robustness of the algorithms	109
6.3 But is it a breakthrough?	112
6.4 Discussion and overall findings	118
6.5 Acknowledgments	124
6.A Supplementary material	125
7 Conclusions	161
7.1 General concluding remarks	161
7.2 Discussion and future prospects	163
7.3 Concluding remarks	177
Bibliography	179
Author Index	195
Samenvatting (Summary in Dutch)	201
Curriculum vitae — Jos Winnink	207

List of Figures

1.1	Front page of United States patent 9,315,437	23
2.1	Number of citations received per year by the key papers and an average HIV/AIDS publication during the first 10 years after publication	32
2.2	Cumulative share of publications as function of the number of citations in the first 12 months after publication	33
2.3	Trend of total number of distinct CWTS subject categories citations originate from	34
2.4	Numbers of ‘new’ subject categories citations originate from added in a year	35
2.5	Trend of scholarly and patent publications in the field of ‘Integrase Inhibitors’	35
2.6	Network of patents and scholarly publications connection the Hazuda paper and the Isentress® patent. Blue (opaque discs) discs represent patent publications; white discs represent scholarly publications	36
2.7	Networks of bridging researchers and their publications related to <i>Isentress</i> ®. Blue (opaque discs) discs represent scholarly publications; white discs represent (co) authors	38
3.1	Graphene related scholarly publications and DOCDB patent families (2000–2012)	50
3.2	Cumulative distribution of graphene related publications from 2004 and 2005 based on the number of citations received during 12, 24, and 36 months after publication. The opaque (red) disks indicate the ‘Novoselov paper’	53
3.3	Scholarly publications citing the ‘Novoselov paper’ disaggregated into science levels	55
3.4	Publications per science category as share of the total number of scholarly publications citing the ‘Novoselov paper’	56

4.1	Numbers of scholarly publications (articles, letters and proceedings papers), patent applications (DOCDB patent families) per year (1975– 2012)	71
4.2	Number of different CWTs subject categories assigned within a year to Ubiquitin-related publications (1980-2010)	72
4.3	Indegree for all publications within the citation network (1951–1986). The dots represent individual publications. The labelled dots represent the five publications with an indegree above 30 .	74
4.4	Citation network of the Ubiquitin discovery based on publications cited by the ‘landmark publications’	75
4.5	Organisational-temporal anatomy of the discovery of Ubiquitin. Shown are publications, (co) authors, research institutions, and interrelations. The opaque (grey) ellipses present the Nobel Prize laureates, and the opaque (grey) disks the ‘landmark publications’. Solid arrows represent affiliation-author relations and citation relations; (co) author-publication relations are represented by dashed arrows	76
4.6	Within-group and out-group publications connected by citation relations with the ‘landmark publications’ (1978-1981)	77
4.7	Knowledge profiles of the Technion Institute (Ciechanover and Hershko) and Fox Chase Cancer Centre (Rose) based on the WoS subject categories of publications up to 1978 covered in the Web of Science database	78
4.8	Cumulative distribution of the number of publications in Biochemical and Biophysical Research Communications published in 1978 versus the number of times cited (1978– 2013)	79
4.9	Within-group and out-group citations for the ‘landmark publications’ (1978-1981) (publication dates)	85
4.10	Within-group and out-group citations for the ‘landmark publications’ (1978-1981) (submission dates)	85
4.11	Publications cited by and citing the ‘landmark publications’ (1950–1986). In Section 4.A.1 on page 84 the meaning of the colors of the discs is explained.	86
5.1	Trends in scholarly publications and patent applications	93
5.2	Frequency counts of unique authors, journals, journal categories, countries, and organisations that produced intron-related scholarly publications	94
5.3	Trends in scholarly publications differentiated by type of science	94
5.4	Term map for scholarly publications published in 1990	95
5.5	Term map for scholarly publications published in 1991	96
5.6	Patent-publication network according to non-patent-reference connections (1980-2010)	96
6.1	Overview of conceptual and analytical framework	104

6.2	Share of papers recognised as breakout as a function of the threshold value applied on Categories	110
6.3	Distribution of the number of citations within 24 months for all papers (articles and letters published in the period 1990–1994), and papers in the datasets Categories, and Clusters	126
6.4	Cumulative distribution for number of papers vs. number of times cited by review papers. Included are all publication (articles, letters) in the WoS published in 1990–1994	127
6.5	Cumulative distribution for number of papers vs. number of times cited by patents. Included are all publication (articles, letters) in the WoS published in 1990–1994	127
6.6	Cross Disciplinary Impact for a sample of seven papers using as measure the number of different ‘WoS subject categories’ of the citing papers at a certain point in time	128
6.7	Cross Disciplinary Impact for a sample of seven papers using as measure the number of different ‘CWTS paper clusters’ of the citing papers at a certain point in time.	129
6.8	Categories: normalized number of papers as a function of the applied threshold. For each variable the number of papers for the threshold value 1 is taken as reference.	130
6.9	Clusters: normalized number of papers as a function of the applied threshold. For each variable the number of papers for the threshold value 1 is taken as reference.	132
6.10	Categories: share of documents identified as breakout as a function of the threshold value	151
6.11	Clusters: share of documents identified as breakout as a function of the threshold value	152
6.12	Categories: difference between the identified number and the expected number of breakout papers as percentage of the expected number of breakout papers	154
6.13	Categories: difference between the identified number and the expected number of breakout papers as percentage of the expected number of breakout papers	156
7.1	Distribution of the number of authors of a paper for ‘breakout’ and for ‘all’ papers from 2007–2011	175

List of Tables

1.1	Cha-Cha-Cha typology of discoveries	11
1.2	Examples of discoveries and their classification by Koshland (2007, p.761)	12
2.1	General description used by De Clercq (2009a) to mark pivotal developments supplemented by the number of citations received by the key papers	31
2.2	Relative position of the key papers within the citation distribution of all HIV/AIDS papers published in the same year*	33
2.3	Short biography of professor Erik De Clercq	40
3.1	Scholarly publications (articles, letters), patent filings and citation to the 'Novoselov paper'	52
3.2	Citations to the 'Noveselov paper' in the first 12, 24, and 36 months after publication, the ranking among all graphene publications from 2004 and 2005, and the ranking among all publications from the same Volume (306) of Science on the basis of the number of citations.	54
3.3	Journal classification science levels	54
3.4	For articles and letters citing the 'Novoselov paper' the quantities <i>NFA</i> , <i>NFO</i> , <i>RelMutNFA</i> , <i>RelMutNFO</i> , <i>NFANFORatio</i> , <i>NFAcitsratio</i> , and <i>NFOcitsratio</i> are compared with the values obtained for the 318 publications from Science Volume 306, and the 66,073 publications published in October 2004. Citations within 24 months after the publication of a cited document are considered.	58
3.5	Information on the professional background of professor Frenken	63

3.6	Evolution of the knowledge base about Graphene using co-occurrence term maps for the period 2005–2012 (Winnink and Tijssen, 2015). The graphs were also presented at the STI conference in 2012 (Winnink, 2012). The maps are based on the abstracts of Graphene related scholarly publications up to the end of the indicated year.	64
4.1	Key figures for the citation network for the period 1951–1986 with (Ciechanover et al., 1978), (Ciechanover et al., 1980b), and (Hershko et al., 1980) as focus points	73
4.2	Key citation figures for the ‘core’ document shown in Figure 4.5. Hershko (1974); Hershko et al. (1978) are conference papers for which no citation information is available.	79
4.3	Based on the number of times cited the position of the four ‘core’ publications from 1980 compared to the publications in the ‘reference set’ is shown.	79
6.1	Robustness of algorithms	110
6.2	Performance of the algorithms on the two datasets	111
6.3	Performance of algorithms within the test set	114
6.4	The citation impact of the breakout papers based on the number of citations received from three different citing sources. The classification is based on the distribution of the numbers of citations received by all papers (letters and articles) published in the period 1990–1994 and covered in the WoS database.	115
6.5	Characteristics of the breakout and non-breakout papers in the test set as a percentage of papers that belong to the top 3% percentile for all papers (articles and letters) based on the number of citations received from the different sources.	115
6.6	Distribution of all 1990–1994 (articles & letters) identified as a breakout paper, along the two dimensions ‘cited by review papers’ and ‘cited by patents’. Based on citation counts from the moment of publication until 2015. (n=1,805,245)	117
6.7	Distribution of all 1990–1994 (articles & letters) cited within 24 months, along the two dimensions ‘cited by review papers’ and ‘cited by patents’. Based on citation counts from the moment of publication until 2015. (n=408,292)	117
6.8	Effect of applying the algorithms to the complete dataset with papers (articles, letters) from 1990–1994 A ‘+’ indicates that share of papers in this category increased, and a ‘-’ that it decreased as result of the selection done by the algorithms.	117
6.9	Key figures for Clusters and Categories	119
6.10	Key figures for the two datasets used	126
6.11	Bibliographic information for the papers mentioned in Figure 6.6 and Figure 6.7	129
6.12	Overlap of the results of Redner’s algorithm and the five algorithms (ARI, DII, RII, CDI, RNI) (no threshold applied)	130

6.13 Categories: number of papers in the Categories dataset, and in the various result sets in relation to the applied threshold	131
6.14 Clusters: number of papers in the dataset and in the various result sets in relation to the applied threshold	131
6.15 Results for the 1990–1994 papers in the top 100 most cited papers of all times (coloured marks indicate that only the CDI_{dc} algorithm recognised the paper in combination with the Clusters dataset)	133
6.16 Nobel Prizes based on scholarly work published in the period 1990–1994; indication is given of whether a paper is recognised and by which algorithm	134
6.17 Papers in the test set, selected on the basis of belonging to the top 10 most cited papers within 24 months for each of the algorithms (an ‘X’ in a cell indicates that the paper is recognised as a breakout paper by the corresponding algorithm; coloured marks indicate that only the CDI_{dc} algorithm recognised the paper in combination with the Clusters dataset)	135
6.18 Validation of the papers mentioned in Table 6.16	139
6.19 Bibliographic information for the papers mentioned in Tables 6.15, 6.16, and 6.17	142
6.20 Key figures for the various algorithms applied to the test set with 60 papers	146
6.22 Number of times a paper in the test set is cited by a review paper	146
6.21 Characteristics of the algorithms by differentiating the papers from the test set in breakout papers and non-breakout papers. Shown are the percentages of papers in each of the two groups that belong to the Top 3% percentile of distribution of the number of citations received from the different sources by all papers (letters and articles) published in 1990–1994 and covered in the WoS database.	147
6.23 Classification of the results based on the total number of citations received in comparison with the distribution of the citation scores for all papers (‘letters’ and ‘articles’) published in the period 1990–1994 and covered in the WoS database; percentile boundaries are presented between parentheses.	150
6.24 Categories: breakout detection specificity (recall rate) of the algorithms. Ratios are based on the absolute numbers in Table 6.13	151
6.25 Clusters: breakout detection specificity (recall rate) of the algorithms. Ratios are based on the absolute numbers in Table 6.14	152
6.26 Detection probabilities for each algorithm on the two datasets .	153
6.27 Categories: expected breakout recall	153
6.28 Categories: difference between the identified number and the expected number of breakout papers as percentage of the expected number of breakout papers.	154
6.29 Clusters: expected breakout recall	155

6.30 Clusters: difference between the identified number and the expected number of breakout papers as percentage of the expected number of breakout papers.	155
6.31 Labels of distribution classes used in tables 6.32–6.40	156
6.32 Categorisation of all 1990–1994 papers (articles, letters) cited within 24 months, along the two dimensions ‘cited by review papers’ and ‘cited by patents’. Based on citation counts from the moment of publication until 2015.	157
6.33 Distribution of all 1990–1994 papers (articles, letters) cited within 24 months, along the two dimensions ‘cited by review papers’ and ‘cited by patents’. Based on citation counts from the moment of publication until 2015.	157
6.34 Categorisation of all 1990–1994 papers (articles, letters) including papers not cited ^a within 24 months, along the two dimensions ‘cited by review papers’ and ‘cited by patents’. Based on citation counts from the moment of publication until 2015.	157
6.35 Distribution of all 1990–1994 papers (articles, letters) including papers not cited ^b within 24 months, along the two dimensions ‘cited by review papers’ and ‘cited by patents’. Based on citation counts from the moment of publication until 2015.	158
6.36 Categorisation of all 1990–1994 papers (articles, letters) identified as a breakout paper, along the two dimensions ‘cited by review papers’ and ‘cited by patents’. Based on citation counts from the moment of publication until 2015.	158
6.37 Distribution of all 1990–1994 papers (articles, letters) identified as a breakout paper, along the two dimensions ‘cited by review papers’ and ‘cited by patents’. Based on citation counts from the moment of publication until 2015.	158
6.38 Effect of applying the algorithms to the complete dataset with papers (articles, letters) from 1990–1994	159
6.39 Categorisation along the two dimensions of all 1990–1994 papers (articles, letters) that are not cited within 24 months, but are cited at least once at a later moment. Based on citation counts from the moment of publication until 2015.	159
6.40 Distribution along the two dimensions of all 1990–1994 papers (articles, letters) that are not cited within 24 months, but are cited at least once at a later moment. Based on citation counts from the moment of publication until 2015.	159
7.1 Distribution of breakout papers (articles, letters) across NOWT categories (2007–2011) ^a	173
7.2 Descriptive statistics for the multidisciplinary journals Nature, Science and PNAS (articles and letters, 2007–2011)	174
7.3 Distribution of breakouts: over and under representation relative to all papers (articles, letters) in the WoS database (2007–2011)	174

LIST OF TABLES

7.4	Percentile borders for the distributions of the numbers of authors for both 'breakout' papers and for 'all' papers from 2007–2011	175
7.5	Weighted average number of author per paper of WoS types article and letter from 2007–2011	175
7.6	Point in time of occurrence of the breakout character of a paper (article, letter), published in the period 1990–1994 vs. organisational collaboration	176

Introduction

1.1 General context

Scientific and scholarly research may result in a new discovery: an observation or finding of something unknown prior to that discovery. The nature and impact of such a discovery on the cognitive structure and evolution of science may vary considerably. Some of those discoveries, each showing a major impact on future scientific research, are considered to signal possible breaches, focus shifts, or even turning points in science. In this thesis the term *breakthrough* is used for those discoveries that have such a major impact on science. Given the vast number of scholarly publications published each year an automated computerised selection system might be a preferable method to harvest databases with bibliographic data of scholarly publications and to search for high-impact publications. Such a generalized and transparent method should facilitate the early and unbiased detection of potentially important new directions in science and technology. An objective method, consisting of one or more algorithms, is relevant as human beings who carry out the evaluation of new developments might be forced to follow a set of strict protocols. The role of these protocols is to prevent preconceptions that could influence this process of evaluation.

The impact of discoveries may extend beyond the domain of science. Scientific discoveries may be crucial steps towards technological applications, and to innovations and products. Several well-known studies searched for the impact of those discoveries on the development of technology¹ and have analysed cases of technological progress (e.g. Globe et al. 1973; Heilbron 1972; IIT 1968, 1969; Isenson 1969; Jewkes et al. 1958; Walsh 1973; Wooding 2007). All these studies indicate that it can take many years before a scientific discovery finds its way into new or adapted technology. This observation is also in line with Grupp and Schmoch (1992) who conclude that the

¹Technology – the application of scientific knowledge for practical purposes.

diffusion of basic scientific knowledge to become the basis of new technology often takes dozens of years² Scientific discoveries and their incorporation in technology are often interlinked in complex ways within research and development (R&D)³ systems, and may span several years, decades, or even centuries.

Forecasting the changes that discoveries may bring about, and monitoring or nowcasting⁴ the evolution of emerging areas in science or technology, presents us with a series of conceptual and methodological challenges. This thesis focuses on how discoveries can change the fabric of science itself, more specifically the immediate impacts within the first two to three years after the discovery was published. Such early-stage detection of major discoveries is relevant not only to scientists themselves, but also to government policy-makers and corporate R&D executives as they may signal significant focus shifts in industrial R&D systems. On the basis of such information funding strategies can be adapted knowing a possible major new development exists. Policy-makers and funding agencies have a particular interest in knowing in which direction research⁵ and innovation⁶ are heading to gain or sustain economic growth and prosperity, or to allocate scarce resources for R&D. These R&D decision-makers usually only oversee the areas of science and technology they focus on, and therefore they might easily miss or misinterpret relevant (fast moving) developments outside their focal area. The newest developments with possible large and immediate impacts on ‘upstream’ science and technology in later ‘downstream’ stages of development are of particular importance.

1.1.1 Research goal and research question

This thesis will focus on one of the upstream stages in the process: new scientific discoveries and the first wave of impacts on their scientific environment. At this early stage, it is usually not clear — neither to the researchers responsible for the discovery nor to external analysts — what the implications are (or could be) in terms of scientific impact or usage outside science. Given their restricted view of R&D decision-makers on these new developments, any source of empirical and objective information that may help provide clarity

²An example is Graphene. Wallace (1947) presented — based on theoretical physical calculations — the properties to be expected for a material currently known as ‘graphene’. It was not until 2004 however, with the publication of Novoselov et al. (2004) when ‘freestanding’ graphene became a reality and the predicted properties could be experimentally verified. The Nobel Prize Physics was awarded in 2010 For to Konstantin Novoselov and André Geim for this discovery.

³R&D — general term for activities in connection with corporate or governmental innovation.

⁴Nowcasting is the activity of estimating the current situation on the basis of historic data.

⁵Research — studious inquiry or examination; especially: investigation or experimentation aimed at the discovery and interpretation of facts, revision of accepted theories or laws in the light of new facts, or practical application of such new or revised theories or laws. Source: <http://www.merriam-webster.com/dictionary/research>

⁶Innovation — the act or process of introducing new ideas, devices, or methods. Source: <http://www.merriam-webster.com/dictionary/innovation>

is highly desirable. To address this information gap, I have developed, tested and applied a series of computerised algorithm-based methods, extracting their information from the vast amount of scientific publications, to identify specific, high-profile discoveries – denoted as ‘breakthroughs’ – that are expected to have a major impact on future research.

The approach used in this study is to analyse, in detail, four case studies of well-known ‘landmark’ discoveries that have been classified as a breakthrough by subject experts. We can benefit from a large pool of information to study each of these well documented ‘exemplar’ discoveries. The bibliographic data related to these discoveries, from the respective research publications, was searched for impact patterns that are characteristic for the early stage evolution of that particular class of breakthroughs. The detection algorithms, designed to single out similar types of breakthroughs, were constructed from those impact patterns.

The objective of this research is to develop and validate methods forming an analytical framework to identify at an early stage high-impact publications that form the basis of new fields in science. An analytical framework should facilitate the nowcasting, and perhaps even the forecasting, of the impact of a scientific discovery on the evolution of science and technology. The founding concepts this analytical framework is based on are (1) a scientific discovery can be considered a discontinuity in the science system as it exists at that moment, (2) and it is possible to detect such discontinuities using bibliographic information. The research goal leads to the following research question:

Is it possible to design, develop, implement, and test an analytical framework and measurement model as a general-purpose tool with a range of practical applications for early detection of breakthroughs in worldwide science?

By focusing on individual scholarly publications this study concentrates on the *micro level* of scientific progress.

1.1.2 Structure of this thesis

This PhD-thesis is structured in the following way. This first chapter continues with sections 1.2–1.5 which introduce (1) the theoretical and conceptual framework of the study, (2) the analytical framework that guides the research design and facilitates data collection, (3) the introduction of the four case studies and the concluding validation study of the detection algorithms, and (4) the information sources used in the study.

The four case studies, which constitute the empirical core of this empirical study, are discussed in full detail in chapters 2–5. Chapter 2 presents the first case study that focuses on the development of ‘Integrase Inhibitors’, a type of HIV/AIDS medicine. The case study analyses the route from the scholarly publication describing the basic biochemical mechanism (Hazuda et al., 2004b) these medicines try to block up to the granted patent that describes

the first drug of this type that was approved for human use. Chapter 3 discusses the second case study, namely the discovery of a method to isolate Graphene⁷, and the effect this discovery (Novoselov et al., 2004) had on the scientific community. The third case study discusses the discovery of the ‘Ubiquitin⁸ mediated proteolytic system’, and is presented in Chapter 4. Particular attention is paid to the fact that this breakthrough was the result of research by a small core team of scholars. Chapter 5 presents case study four that discusses the follow up of the discovery of ‘Introns⁹’ in 1977 by Phillip Allen Sharp and Richard J. Roberts. The focus in this case study is on the fact that after the first breakthrough discovery a second breakthrough took place. This second breakthrough is marked by a shift from discovery-oriented research to more application-oriented research. Chapter 6 presents the study that validates the five algorithms that arose from the four case studies and can be applied to raw bibliographic data without the need of extensive pre-processing to clean and harmonise the data. Chapter 7 synthesizes the empirical work, introduces an alternative definition for a breakthrough, discusses the results, points to possible directions for future research, presents some preliminary results of my follow-up research, and ends with a final remark.

1.2 Theoretical and conceptual framework

1.2.1 Science as a dynamic system

Science can be considered a dynamic system¹⁰ in which scholars and their research activities play a dominant role, and discoveries can act as events that change the nature, shape or direction of scientific progress – either in terms of new knowledge production, or an interpretation or reinterpretation of existing knowledge, ideas and know-how. In general, systems operate in the vicinity of a certain equilibrium state and are considered to be stable unless factors force the system to undergo larger-than-usual changes and to enter a state from which it cannot readily return to the previous situation. Such changes from one stable state to another are also called ‘phase transition’ or ‘phase change’, in analogy to comparable processes in physics, chemistry and biology. The dynamic behaviour of complex systems is covered extensively in the scholarly literature for instance in (von Bertalanffy, 1969). All systems, not just the large ones, can undergo irreversible¹¹ changes (Mandelbrot, 1982). Several empirical studies have shown that dynamic systems

⁷Graphene is a single, tightly packed layer of carbon atoms that are bonded together in a hexagonal honeycomb lattice. Graphene has a number of extraordinary properties.

⁸Ubiquitin is a label-molecule that is attached to a protein that is to be destroyed. Shortly before the protein is squeezed into the proteasome, its ubiquitin label is disconnected for re-use.

⁹A non-coding, intervening sequence of DNA within a gene

¹⁰System – a group of related parts that move or work together(<http://www.merriam-webster.com/dictionary/system>).

¹¹Without external influence the system is incapable of returning to the previous condition or state.

1.2 Theoretical and conceptual framework

can transmit early-warning signals indicating a ‘phase transition’ is about to happen Lade and Gross (2012); Scheffer (2009, 2010); Scheffer et al. (2009); a transition to a new state in which it stays until a new event forces the system to move to yet another state. Such a major transition stands out between the more common ‘minor’ changes a system undergoes frequently. Whether or not a scientific discovery should be considered a minor or major change — a breakthrough or not — has been a topic of study and academic debate during the last 50 years.

1.2.2 Progress in science

There is a general notion that science progresses on the basis of work done by scholars, researchers and scientists that builds on prior achievements (often by others); as described by the motto “If I have seen further it is by standing on the shoulders of giants”¹². The evolution of science, however, does not follow a linear, continuous, cumulative unified path, which is the impression of the development of science as it emerges from textbooks (Kuhn, 1962), where the knowledge is ordered in such a way that it can serve education. Kuhn distinguishes ‘normal’ science and ‘revolutionary’ science, and argues that the development of science alternates between these two states. In normal science discoveries fit within an existing paradigm¹³ and are expected¹⁴. Revolutionary science deals with those discoveries that are at odds with the then existing paradigm.

Normal science, in Kuhnian terminology, is scientific research conducted within a single paradigm. Within normal science the foundations of the paradigms and the paradigms themselves are not argued, and science research functions as a ‘puzzle-solving’ activity inside a framework of common understandings and starting points. At the point when *tension* between the then current paradigm and observations from scientific research occurs, a new paradigm might come into existence, in which case a ‘paradigm shift’ can be observed. Wray (2011, p.202) argues: “...According to Kuhn’s mature view, a new theory is developed in a field in an effort to account for an anomaly that the accepted theory was unfit to account for ...”. The new paradigm enables the resolution of previously unsolvable problems and replaces the old one. Paradigm shifts proliferate slowly as the relevant scientific community needs to be convinced to alter its views and approaches. This process will go on forever. Kuhn’s observation of discontinuities in the development of science is now widely accepted. Results of scientific research that can only be explained by changing an existing paradigm are characteristic for revolutionary science, according to Kuhn. Radical novel approaches, new information

¹²This metaphor is usually attributed to Sir Isaac Newton, but should be ascribed to Bernard of Chartres as it was first recorded in the 12th century (Merton, 1965, p.267)

¹³Kuhn (1962) defines a paradigm as “...that which the members of a scientific community, and they only share ...”

¹⁴Although there is a sense that a discovery is forthcoming the exact moment it will happen is uncertain.

and discoveries, which are incompatible with the current dominant theoretical framework and beliefs within a science field, may suddenly appear on the scene and revolutionize the cognitive structure of that field (Andersen et al., 2006). These are the ‘phase transitions’ that have a large impact on science within a relatively brief span of time.

1.2.3 Discoveries in science

Identical or related discoveries frequently come in a manifold - “...It is an interesting phenomenon that many inventions¹⁵ have been made two or more times by different inventors, each working without knowledge of the other's research ...” (Ogburn and Thomas, 1922). Such ‘multiple discoveries’ may differ in appearance, and occur at different points in time, or at different geographical locations. Merton (1961, Ch.II, p.478), who confirms the observations made by Ogburn and Thomas (1922) expands on the notion of manifold discoveries in the following way:

“...These seeming singletons¹⁶ which then turn out to have been multiples or rediscoveries. Other, more compelling, classes of evidence bear upon the apparently incorrigible hypothesis that singletons, rather than multiples, are the exception requiring distinctive explanation and that discoveries in science are, in principle, potential multiples ...”

Merton (1961, p.480) also refers to his study on historical incidents of multiple discoveries in which he reported on the occurrence of up to five and six-fold discoveries. Price (1963, p.65–66) also discusses this phenomenon and links it with Kuhn's concept of normal science in which discoveries in a sense are to be ‘expected’ from time to time. Simonton (1978, 1979) and Brannigan and Wanner (1983) analysed historic data on sequences of discoveries in science to uncover the mechanism behind the phenomenon of multiple discoveries. Brannigan and Wanner (1983) conclude that of the several possible stochastic models that can be used to describe the distribution of the grade of multiples, models based on a Poisson distribution gives adequate results.

1.2.4 Scholarly communication

Scientists generally use the results achieved by other researchers as a starting point for their scientific insights, as is expressed by the already mentioned metaphor *“If I have seen farther, it is by standing on the shoulders of giants”*. A discovery can only contribute to advances in science when it is codified in a way that allows communication to others. The principal means of scholarly communication is by text; in modern-day times usually by research articles

¹⁵Nowadays the terms ‘discovery’ and ‘invention’ have distinctive and separate meanings. In the past these terms were used interchangeably, and ‘invention’ was also used in situations where currently the term ‘discovery’ is preferred.

¹⁶Singletons – discoveries occurring only once.

1.2 Theoretical and conceptual framework

in scholarly or technical journals ('research publications') or in books. Price (1963, Ch.3, p.68) discusses the role of scientific publications, and concludes:

"... The scientific paper therefore seems to arise out of the claim staking brought on by so much overlapping endeavour. The social origin is the desire of each man to record his claim and to reserve it for himself ..."

Sharing and claiming research findings is a major reason for communication within the scientific community. Communication in science takes place in several forms. Formal means of communication are scholarly publications, conference proceedings, and books. Less formal¹⁷ ways of communication play a role within teams of collaborating researchers who have close working relations in which information sharing is obligatory for the team to be able to function optimally. Crane (1972) concludes, that as scientists rely on research results of other researchers, they group together in 'invisible colleges'. The citing-cited relations between scholarly publications form the fabric of these invisible colleges. Formal means of communication within the virtual colleges are publications (Lievrouw, 1989). Price (1965) argues that the pattern of bibliographic references reflects the nature of the scientific research front. Lievrouw (1989, p.616) examines the relationship of bibliometric techniques, especially citation analysis, with communication theory and research, and argues "... However, it¹⁸ is of particular interest here because it is possibly the best-known model of scientific communication ...".

Citation analysis – e.g. (Moed et al., 2004) – acts as an important framework for the analysis of various aspects of the scientific community, and is a central theme in bibliometrics¹⁹. Developments in science can be monitored using citation relations between scholarly publications. Citation relations between patent publications and scholarly publications provide an – albeit partial – view on the influence of scientific research on technological evolution. Citations are biased in the sense that they are influenced by several mechanisms that are not directly related to the contents of the publication. Price (1963, p.87) mentions that in certain situations where the results of team research are reported "... The participating physicists are not mentioned, not even in a footnote ...". Merton (1968) points to psychosocial conditions that have an impact on citation behaviour, for instance already eminent researchers are given disproportionate credit in some cases. Crane (1972, p.83) concludes that social factors within a research field have an effect on the diffusion of knowledge in the field, and that these factors furthermore determine which information is to be used in later publications. Notwith-

¹⁷Data from social media is not taken into account in this study as this is (1) a recent development and (2) the value for the analysis such as those carried out in this study is not yet evident.

¹⁸[Added by the author] with 'it' Lievrouw refers to the concept of 'invisible colleges'

¹⁹Pritchard (1969, p.349) defines 'bibliometrics' as "... the application of mathematics and statistical methods to books and other media of communication ...", and by Broadus (1987, p.376) as "... the quantitative study of physical published units, or of bibliographic units, or of the surrogates for either ...".

standing the limitations, citation relations can be used as a proxy to reveal the evolution of science and technology.

The bibliographic information of scholarly publications is used in this study as a major information source. These publications do not form a homogeneous group as they present various forms of dissemination of information between scholars. A document type is assigned to each scholarly publication when the accompanying bibliographic information is stored in a database. One of the assigned classes is ‘article’. Articles are considered to be the publications that contain the results of original scientific research. These publications are often multipage publications published in a scientific journal. The concept of an article seems obvious at first sight, but consists of several types of publications and contains multipage publications as well as shorter publications known as, for instance, ‘letters to editor’ or ‘opinion letters’. Such shorter publications may also contain the results of original research or can contain original ideas²⁰.

1.2.5 Discoveries and breakthroughs

Scientific practice can be seen as a system that continuously undergoes changes as a result of discoveries that influence the science system. Every ‘open’ scientific discovery - one that is properly documented and communicated to others within the relevant research community - is likely to have an impact, although (at first) possibly negligible, on (r)evolutionary changes in science. The discoverers and other subject experts are able, usually with the benefit of hindsight, to identify and value those impacts after a period of time.

When studying the evolution of science fields, the flow of publications is often used as an approximation of the dissemination of the knowledge related to a scientific finding and of the way other researchers follow-up on this finding. Knowledge diffusion does not guarantee a continuous gradual evolution of science because the diffusion process changes parameters in the system and can therefore result in unexpected high-impact changes. Some discoveries might not be noticed for some time, for a many reasons.²¹ Some might even be totally neglected where the result is not properly documented or communicated - the result is forgotten or its implications overlooked. Only a small number of scientific discoveries lead to large, structural changes in science fields, and pave the way for novel insights and further productive research. For a discovery to be qualified as a distinctive ‘major’ discovery not only requires the judgement by a wider range of subject experts, but also needs sufficient length of time to allow extensive validation and general appreciation. Becattini et al. (2014)²² analysed the time lag between discoveries

²⁰Koshland (2007) is an example of a two-page manuscript typified as ‘opinion letter’ containing relevant original ideas to which the document class ‘article’ is assigned.

²¹The fact that a discovery might remain unnoticed by the scientific community for some time, but is later considered a breakthrough, is for instance described by Ciechanover (Ciechanover, 2009, p.2)

²²Becattini et al. (2014) “...After 1985 about 15% of physics, 18% of chemistry, and 9% of medicine prizes were awarded within 10 years of the corresponding discoveries. By contrast,

1.2 Theoretical and conceptual framework

and the awarding of a Nobel Prize and conclude that Nobel prizes are awarded only very rarely within 10 years of a discovery.

The term ‘breakthrough’ is usually applied to such discoveries, a term frequently used for events that are considered major discoveries. Major journals like *National Geographic*, *Nature* and *Science* regularly publish overviews of what they regard as the major scientific discoveries in a previous period or specific year; these lists are usually based on expert opinions. What exactly is meant by a breakthrough or major discovery is not specified, and for good reason: there is no generally accepted, let alone a universal, definition that can count on full support throughout the scientific community. In particular the notion “... new way of thinking about a problem ...” is an essential property of a breakthrough put forward by Hollingsworth (2008, p.317). The term breakthrough is not only used for the nature of the transition (What exactly did change?) but is also used for the point in time the event occurred (When did the change occur?) or to the impact the discovery had on other systems. The fact that the same term is used for the closely related phenomena is elucidated in Hofstadter and Sander (2013, Ch.1).

This absence of a generally accepted definition is illustrated by the fact that various synonyms are in use for the term breakthrough such as ‘advance’, ‘development’, ‘step forward’, ‘quantum leap’, ‘evolution’, and others. The lack of a single definition for a breakthrough complicates the identification of these phenomena. Hollingsworth (2008, p.317) defines a breakthrough as “... A major breakthrough or discovery is a finding or process, often preceded by numerous small advances, which leads to a new way of thinking about a problem ... This new way of thinking is highly useful to numerous scientists in addressing problems in diverse fields of science ...”. Hollingsworth argues further that science evolves not just through the occasional breakthroughs but also by means of numerous successive small, incremental advances. The co-existence and interplay between ‘incremental’ and ‘breakthrough’ advances is in line with Kuhn’s idea that after a ‘paradigm shift’ (i.e. revolutionary science) has occurred ‘normal science’ will take over – at least for some period of time (Kuhn, 1962).

In spite of the lack of proper operationalization and identification, there is general agreement only on the fact that breakthroughs are, by definition, rare events. The precise moment such a major change occurred is even in retrospect hard to pinpoint, and foretelling when such an event is likely to occur is near possible. Nonetheless, some progress is currently being made on theoretical and empirical models that may enable forecasting or prediction methods. For instance, Ball (2004) discusses the fact that systems need a certain ‘critical mass’ to undergo a major change. Complex dynamic systems can have ‘tipping points’ (Scheffer, 2010; Scheffer et al., 2009). The prediction of such tipping points before they are reached is, however, extremely difficult. Scheffer et al. (2009) concludes “...work in different scientific fields is now suggesting the existence of generic early-warning signals that may indicate

before 1940 about 61% of physics, 48% of chemistry, and 45% of medicine prizes were awarded within 10 years of the corresponding discoveries ...”

for a wide class of systems whether a critical threshold is approaching ...”.

Breakthrough discoveries are events that have a major impact on future scientific research and can be considered a tipping point in science. Several scholars constructed theoretical models that focus on the diffusion of knowledge within the scientific community, and connect this knowledge diffusion with the occurrence of discoveries. Andersen et al. (2006) focus on the cognitive changes that occur in science when a paradigm shift occurs. In the publications of Bettencourt and colleagues (Bettencourt and Kaiser, 2011, 2015; Bettencourt et al., 2009) a *percolation model* describing the development of science is proposed. Bonaccorsi (2008, 2010) hypothesises that new science fields that came into existence after the 1970s follow a different evolutionary development path compared to already established sciences. Chen et al. (2009) propose an explanatory and computational theory of transformative discoveries in science. Cintron-Arias et al. (2005) tested mean-field deterministic epidemic models to describe knowledge diffusion. A catastrophe model to develop a formal nonlinear model of scientific change in concordance with Kuhn’s hypotheses is put forward by Perla and Carifio (2005). Sung (2008) shows that experiments play a crucial role in formulating an explanation of the RNAi anomaly. Vitanov and Ausloos (2012) focus on the uses of compartmental epidemic models²³, Lottka-Volterra’s model and others to describe technology diffusion. These and other theoretical models contain conceptual descriptions of the evolution of the science system, and can therefore be used to identify areas that evolve into hot spots.

1.2.6 Characterizing discoveries

Time not only picks the winners, it also unmaskes discoveries that turned out to be hypes²⁴, hoaxes and frauds. An example of a hoax is the claim for the existence of nuclear fusion at room temperature – ‘cold fusion’ (Fleischmann and Pons, 1989). This claim was almost immediately criticized, and it was concluded (Dmitriyeva et al., 2012) that “...According to our calculations, the experimentally measured excess heat can be accounted for fully by this chemical reaction ...”. The Korean researcher Hwang Woo-Suk was considered one of the pioneering experts in the field of stem cell research until a publication by Cyranoski (2004) uncovered Hwang’s fraudulent research. The increasing incidence of retraction of scientific publications (Cokol et al., 2008) blurs the picture of the bibliographic data. Clearly the term ‘breakthrough discovery’ should be used with caution. So, perhaps it is not surprising that in the academic literature there is a lack of convincing typologies or helpful classification systems of scientific discoveries. One of the exceptions is the ‘Cha-Cha-Cha’ theory developed by Koshland (2007), who classifies scientific discoveries into three distinct classes based on the nature of the discovery

²³Compartmental epidemiological models with a name based on the specific compartment structure of the model, e.g. SIS, SIR, SEIR.

²⁴Hype – extravagant or intensive publicity or promotion (source: Oxford Dictionary of English, 2nd edition)

1.2 Theoretical and conceptual framework

Table 1.1: Cha-Cha-Cha typology of discoveries

Koshland type	Type of discovery	Kuhnian type	Characterisation
Charge	These discoveries solve problems that are quite obvious, but in which the way to solve the problem is not so clear	<i>Normal science</i>	‘...In these discoveries, the scientist is called on, as Nobel laureate Albert Szent-Györgyi put is “to see what everyone else has seen and think what no one else has thought before ...”’(Koshland, 2007, p.761)
Challenge	These discoveries are a response to an accumulation of facts or concepts that are unexplained by or incongruous with scientific theories of the time.	<i>Revolutionary science</i>	“Sometimes the discoverer sees the anomalies and also provides the solution. Sometimes many people perceive the anomalies, but they wait for the discoverer to provide a new concept.”(Koshland, 2007, p.761)
Chance	These discoveries are often called serendipitous	<i>Revolutionary science</i>	‘finding the unsought’ (van Andel, 1994), like the discoveries of penicillin or Teflon® (Koshland, 2007)

in relation to already existing scientific knowledge: **Charge**, **Challenge** and **Chance**. Koshland’s (2007) classification, focusing on how a discovery is different from the then existing scientific knowledge, is just one way of classifying discoveries. Scientific discoveries can be classified on the basis of various characteristics. Redner (2005) for instance classifies discoveries that are documented and presented in a scholarly publication, according to citations of those publications by other scholars. We will return to Redner’s interesting approach, which classifies discoveries as either *non-breakthrough* or a *breakthrough*, in Section 1.3.1 Table 1.1 presents key Cha-Cha-Cha characteristics while Table 1.2 presents some illustrative examples selected by Koshland. Based on the nature of a discovery, events are categorized into one of the three Cha-Cha-Cha types. Discoveries of type Charge are the most common.

1.2.7 Collaboration and research teams

Discoveries are not only about those by individual, prize-winning ‘giant’ researchers and scholars, those who allegedly “stand on the shoulders” of other preceding giants, they are also about joint efforts and research collaboration between individuals benefiting from each other’s knowledge, inspiration and know-how. Three of the discoveries listed in Table 1.2 reflect such ‘team work’, research where two or more researchers played a key role in the working leading up to the breakthrough. Watson & Crick’s discovery of DNA in the 1950s is probably the most well known example in contemporary science. As modern science itself has become much more collaborative (Wuchty et al., 2007), especially ‘big science’ based on large shared research facilities (Price, 1963), the same is likely to apply to scientific breakthroughs. In an empirical

Table 1.2: Examples of discoveries and their classification by Koshland (2007, p.761)

Problem that needed solving	Discovery	Discoverer	Category of discovery
Movement of stars, Earth, and Sun	Gravity	Newton	Charge
Preventing heart attacks	Cholesterol metabolism	Brown & Goldstein	Charge
Why offspring look like their parents	Laws of heredity	Mendel	Charge
Structure of C ₆ H ₆	Benzene structure	Kekulé	Challenge
Constant speed of light	Special relativity	Einstein	Challenge
Atomic spectra that could not be explained	Quantum mechanical atom	Bohr	Challenge
How DNA replicates and passes on coding	Base pairing in double helix	Watson & Crick	Challenge
Clear spots on Petri dish	Penicillin	Fleming	Chance
Preventing heart attacks	Cholesterol metabolism	Brown & Goldstein	Charge
Crystals of D- and L-tartaric acid	Optical activity	Pasteur	Chance
Reagent 'stuck' in storage cylinder	Teflon®	Plunkett	Chance

study Uzzi et al. (2013) conclude that teams are 37.7% more likely than solo authors to insert novel combinations of prior work into familiar knowledge domains. Publications with such novel and unusual combinations are rare but are twice as likely to become highly cited works.

Analysing team collaboration in general one of the main conclusions drawn by Uzzi and Spiro (2005, p.492) is "... Small world networks²⁵ do benefit performance but only up to a threshold, after which the positive effects of small worlds reverse ...". Whitfield (2008, p.720-723) concludes that research is becoming more and more a matter of team activity and the contribution of single authors to science is dwindling. Green and Brendsel (2008) respond that this observation might be correct, but "... Lightning can still strike the solitary explorer whose mind is prepared. ...". These results are in line with Wuchty et al. (2007) who conclude that knowledge-producing teams increasingly dominate over solo authors, and that their publications are more frequently cited; this citation advantage increases over time. These authors furthermore conclude that the process of knowledge creation has undergone a fundamental change in moving from research conducted by individual researchers to research carried out by teams of researchers.

²⁵[Added by the author] A small-world network is a type of mathematical graph in which most nodes are not neighbours of one another, but most nodes can be reached from every other node by a small number of hops or steps. Source: [https://en.wikipedia.org/wiki/Graph_\(discrete_mathematics\)](https://en.wikipedia.org/wiki/Graph_(discrete_mathematics))

Burt (2004) concludes that the opinion and behaviour of people belonging to a group are more homogeneous within a group than between groups. Individuals who are part of multiple groups can therefore bridge cognitive gaps between groups and come up with solutions that otherwise might be unseen. Guimerà et al. (2005) conclude that the forming of a large group of practitioners can be described as a 'phase transition' after having analysed more than 4 million publications issued in a period of more than 30 years. Jones et al. (2008, p.1261) conclude that collaboration in science is increasingly becoming composed of co-operations spanning university boundaries. According to Skilton (2009) articles co-authored by teams that include frequently cited scholars and teams whose members have diverse disciplinary backgrounds are cited more often. Weinberg (1970, p.1056) argues that the formation of large interdisciplinary teams centred on pieces of expensive equipment causes the increasing importance of team science, especially since World War II. Such research teams are said to be part of 'big science' (Price, 1963). Workgroups construct a common group identity over time through the process of value convergence between group members (Meeussen et al., 2014). Bettencourt et al. (2008) analyse the quantitative social structures of collaboration that develop as new scientific fields emerge. An increased interaction between scientists exploring different aspects of a problem creates new concepts, techniques and a shared research programs resulting in successful new fields

Research teams are not fixed structures and evolve over time. Tuckman (1965) introduces what has become known as the standard model of small group development in which four stages are distinguished. McGrew et al. (1999) extend Tuckman's model with three declining phases to create a model that describes both the formation and the decay of small groups. Bettencourt and Kaiser (2011) point to the difficulty in defining and comparing science fields, and observe that there seems to be a general sense that the different fields undergo similar stages of development. Of particular importance in a field's history are the moments at which conceptual and technical unification allows the widespread exchange of ideas and collaboration. These moments mark the point in time when the networks of collaboration between scholars show the analogue of a percolation phenomenon, and develop a giant connected component containing most authors of a conceptual framework.

1.3 Analytical framework

1.3.1 Facing the methodological challenge

Bibliographic information within a research publication that first describes a discovery may refer to its relevance or anticipated relevance - for scientific progress, but its true impact depends on its reception and implementation over time. Subsequent research publications, referring to the discovery and its list of literature references ('citations'), will reflect and reveal reactions from peers in the scientific community to the breakthrough work. As a consequence bibliographic information can only help identify breakthroughs in

those cases where these scholarly publications receive exceptional scores on citation-impact metrics²⁶. Publications with a large ‘citation impact’, or those that are immediately cited, are more likely to be seen - with hindsight - as breakthrough publications by the scientific community. But reaching such shared opinion takes time - often many years or decades. In this thesis, we will refer to highly cited publications that have not (yet) acquired breakthrough status as *breakout* publications, or *breakthrough by proxy*²⁷. Only after sufficient time has elapsed, and with the benefit of expert opinions, will some breakouts be considered a breakthrough. As put forward in Section 1.1.1 on page 2 this study addresses the following methodological challenge:

“Is it possible to design, develop, implement, and test an analytical framework and measurement model as a general-purpose tool with a range of practical applications for early detection of breakthroughs in worldwide science?”

In a first step towards operationalization²⁸, we will focus our attention on identifying breakout publications that are characterised by specific citation impact profiles. To do so, this study focuses on the observable effects of discoveries on the research community, guided by the research question:

“What kind of detectable evidence do discoveries leave behind in the research literature in terms of sudden changes and distinctive structural developments?”

Bettencourt et al. (2009, p.220) hypothesize “... there is a universal character in discoveries ...”, and argue that circumstantial evidence for the existence of such universal characteristics is also supported by several other studies, such as Gerstein and Douglas (2007); Leskovec et al. (2005); Uzzi and Spiro (2005). Chen et al. (2009) introduce an explanatory and computational theory of ‘transformative discoveries’ science based on the central premise of the connection of disparate areas of knowledge is introduced. Their theory explains the nature of these discoveries, and also characterizes the subsequent diffusion process. According to the authors the primary value of the theory is that it provides both a computational model of intellectual growth, and concrete and constructive explanations of where insightful inspirations for transformative scientific discoveries can be found.

Tapping into universal characteristics therefore opens up the possibility of designing early-detection models of breakouts and breakthroughs, either models based on small-scale case studies or those derived from large-scale quantitative analysis. Julius et al. (1977) is an early example of the former, using expert knowledge only. The aim of this thesis is to tackle this question systematically and in a large-scale ‘macro-level’ fashion, i.e. scanning world

²⁶Metric – a technical system or standard of measurement.

²⁷This group is further broken down into subcategories as explained on in Section 6.A.9 on Page 156

²⁸Operationalization is the process of strictly defining variables into measurable factors.

science for breakouts. Obviously, one cannot rely on ‘micro-level’ individual expert judgements (or expert panels) to identify and check each and every of the hundreds or thousands of potential breakthroughs. External observers and analysts, who are not experts in the field under study, will have to resort to other information sources — notably the citations between research publications. Of course, these citations are also expert based, albeit indirectly: each citation from a fellow scientists or scholar to that specific publication describing the discovery can be seen as a ‘vote of relevance’, an expert-based confirmation that the cited work has been noted or had an impact on follow-up scientific research.

The key methodological challenge is to develop citation-based early-detection algorithms that enable large-scale scanning of the global scientific literature — computerised algorithms to identify at an early stage those scientific publications that have, or are likely to have, an above average impact on science. For a computational perspective, an expanding set of citation-based algorithms all build on earlier research methodologies aimed at identifying and monitoring ‘emerging topics’, ‘emerging technologies’ or ‘research fronts’ in scientific progress, technological development, or R&D-based innovations. These ‘tracing and tracking’ methods do not focus on individual publications, but rather on large sets of published documents (usually research publications and/or patents); they also tend to adopt longer time periods. The US researcher Henry Small pioneered the large-scale analytical approach in the 1970s. He introduced the notion that rapid shifts in research focus, as identified in the scholarly research literature, could be regarded as a signal of ‘revolutionary’ change (Small, 1977). More than 30 years later, his research program is still on-going — Henry Small and his colleagues combine direct citations and co-citations helps to adequately identify emerging topics (Small et al., 2013).

The citation-based algorithms introduced in this thesis are closely related to work by Redner (2005) who classified discoveries based on the number of citations of the publications. More recently, Baumgartner and Leydesdorff (2014) applied ‘group-based trajectory modelling’ to citation curves of research publications. Ponomarev et al. (2012) focus on citation patterns in combination with statistical modelling, while a follow-up study by Ponomarev et al. (2014b) focuses on the effects of interdisciplinarity in the subject categories, and geographical diversity. Schneider and Costas (2015) also use citation-based methods to detect potential breakthrough publications. Wang et al. (2016) introduce a measure for the combinatorial novelty of a paper to identify those that are likely to have an above average or even high-impact. The approach adopted in this thesis differs significantly as it not only takes as its leading principle the number of citations a publication receives, but also focuses on the dynamic influence a publication has on the scientific community. This dynamic influence is expressed not only in the number of citations but also in the sources of the citations, like for instance authors, science fields, and clustering of citations. Another difference is the focus on the period from the publication of a paper until three years later.

1.3.2 Designing the early-stage detection algorithms

The research in this study is, as mentioned earlier, rooted in the assumption that bibliographic information for scholarly publications can be used as a proxy to analyse and monitor (sudden) developments in science. Citation links between publications form the basis for measurement. These citation links form the citation profile of a publication that varies over time as a publication gets more and more citations. The response of the scientific society on a publication is reflected in the number of times a publication is cited, and reflects the impact a publication has on the evolution of science, as far as it can be decided on the basis of citation patterns. A basic assumption is that researchers working in the same area are able to value a discovery in relation to already existing knowledge. The impact of a publication on science is in this way linked to the way it is cited²⁹. As a consequence of this approach informal communication that might take place is not taken into account.

Contrary to earlier work and the methods briefly described above, the focus in this study is on the citation impact behaviour of individual scholarly publications relatively soon after publication. The detection method covers a range of citation-based criteria, which are identified by studying the seminal research publications of generally acknowledged breakthroughs. The underlying distinctive citation impact patterns of these 'breakthrough exemplars' are unravelled and their key characteristics are used as a 'citation profile' to design computerised algorithms for searching *Challenge* and *Charge* types of breakthroughs in the global research literature. These early detection algorithms should be able to: (1) identify research publications describing discoveries that are now regarded as breakthroughs, (2) track down 'breakout' discoveries that have not yet been recognized as such.

This approach relies on systematic large-scale searches within the worldwide scholarly literature. Assembling information from large, international bibliographic databases enables external, independent analysis to identify significant short-term³⁰ changes in publication and citation patterns. In this study, the bibliographic is extracted from the *Web of Science Core Collection* database³¹ (abbreviated here to WoS). Further information about this information source, and relevant measurement details, are provided in Section 1.5. The analytical procedure is divided into the following seven steps:

1. Search for characteristic citation patterns of the selected breakthrough publications;
2. Selection of distinctive patterns to construct and test the search algorithms;
3. Selection of WoS-indexed research publications in the period 1990–1994

²⁹In this study the citation information is used 'as is'.

³⁰In this study 'short-term' refers to the period 2-3 years immediately after publication of a research paper (indexed by the WoS) in which the discovery is first introduced and/or described.

³¹Web of Science Core Collection is a product of Thomson Reuters, the USA-based media company. The division of Thomson Reuters responsible for the WoS has recently (July 2016) been sold to two investment firms: Onex Corporation and Baring Private Equity Asia.

- (focussing original research findings published in ‘article’, and ‘letter’ document types);
4. Determining ‘optimal’ citation frequency threshold values to pre-select cited publications;
 5. Construction of two datasets with WoS-indexed publications published in 1990–1994, with publications that belong the top 10% most cited within two years after publication. These sets are based on two types of research subfields: (1) WoS-related *subject categories*³² (2) in-house defined *document clusters*³³;
 6. Application of all the developed detection algorithms to both datasets;
 7. Quantitative, statistical analysis of the results.

The findings from step 7 are presented in a series of research articles on each case study, supplemented by a concluding ‘validation study’. The next section introduces these empirical studies.

1.4 Introduction of the studies in this thesis

1.4.1 Criteria used to select the cases to be studied

Five separate studies were undertaken, four of which are case studies while the fifth, and concluding study validates selected results from those case studies. The general criteria for a subject of a case study to be included in this study are (1) the discovery is characterised by the scientific community as a ‘major step forward’, i.e., a generally recognized ‘breakthrough’; (2) information about the seminal publication and publications citing it should be covered by the available data sources; (3) the discoveries have led to technology as evidenced by the occurrence of patent publications that can be linked with the discovery; (4) the time lag between the publication of the seminal publication and the moment of analysis should be large enough to allow for patents publications to be available; (5) the case studies must be examples of ‘charge’ and ‘challenge’ discoveries³⁴; and (6) the number of case studies is restricted to four. In Section 1.3.1 it is argued that a universal character in discoveries and inventions exists. To uncover this universal character a small number of case studies should be sufficient. Four case studies are considered to uncover enough characteristic patterns to answer the research question put forward in this study.

The four selected case studies belong to two different ‘Koshland types’. One group contains ‘Charge’ breakthroughs and the second group ‘Challenge’ breakthroughs. The study on Integrase Inhibitors (Section 1.4.2) – a HIV/AIDS medicine – and the study on Graphene (Section 1.4.3) make up this first group.

³²In the WoS 251 different subject categories describing different fields in science are defined.

³³A document classification method based on citation relations between publications is developed (Waltman and van Eck, 2012) as an alternative for the WoS subject categories.

³⁴Discoveries of the ‘chance’ type are not taken into account, as they are very rare as explained in Section 1.2.6

Both breakthroughs are examples of ‘charge’ breakthroughs. The case studies in the second group focus on two ‘Challenge’ breakthrough discoveries: (1) the discovery of ‘Ubiquitin-mediated proteolytic system’ (Section 1.4.4), and (2) the discovery of ‘Introns’ (Section 1.4.5). For these well-documented discoveries practically all relevant-information, in retrospect, is available. Discoveries of type ‘Chance’ are so rare, and supposed to be unpredictable that they are not included in this study. The study to validate the algorithms that are constructed on the basis of the results of the analyses of the four case studies is introduced in Section 1.4.6.

1.4.2 Case study #1: Integrase Inhibitors (Chapter 2 of this thesis)

In medicine, it often takes 10 to 14 years from ‘upstream’ disease characterization to ‘downstream’ drug approval for human use. In the 1980’s a novel method for drug design, known as ‘rational drug design’ (RDD), replaced the then existing trial-and-error harvesting methods for finding new lead compounds that could act as a basis for a new medicine. The RDD success encouraged the development of various computational methods that use structural information for suggesting novel molecular structures, and this method has led to a drastic reduction in the development time of a new drug. Using the RDD route, the speed of the development of medicines increased tremendously. In this case study the development of ‘Integrase Inhibitors’, a type of HIV/AIDS medicine, is analysed. This study, which is presented in Chapter 2, focuses in particular on links between science and technology embodied in scientists who also act as inventors.

1.4.3 Case study #2: Graphene (Chapter 3 of this thesis)

The Nobel Prize for Physics 2010 was awarded to André Geim and Konstantin Novoselov for “succeeding in producing, isolating, identifying and characterizing graphene” (Nobel Prize Committee, 2010). Graphene seems to be an appropriate case because the publication in which the discovery was published Novoselov et al. (2004) acted as a bridge between basic scientific research and applied science, and thus marks the beginning of technological applications. Prior to this seminal publication graphene was only studied theoretically and experimentally as an integral part of graphite crystals that, as a whole, do not have the physical properties of freestanding Graphene. We discuss this case in Chapter 3, focusing on the inflow of researchers and research institutes into the field of Graphene research.

1.4.4 Case study #3: Ubiquitin (Chapter 4 of this thesis)

The original discovery of the ‘Ubiquitin-mediated proteolytic system’ was published in (Ciechanover et al., 1978) and was the result of the research efforts of a small team of scientists. For some years the discovery remained unnoticed

1.5 Information sources

(Ciechanover, 2006) by researchers outside the small research group. The Nobel Prize in Chemistry was awarded for this discovery in 2004. This discovery is an example of a ‘Challenge’ type discovery and led to many R&D activities. The Ubiquitin case is an example of a discovery that in the beginning was carried by a small research team and is discussed in Chapter 4.

1.4.5 Case study #4: Introns (Chapter 5 of this thesis)

The scientific discoveries underpinning this research were published in 1977 and revealed that chromosomes comprise of a mosaic of ‘exons’, used to form new copies of genes, and of ‘introns’ that appeared to be non-functional ‘interspaces’. These discoveries proved the assumption ‘chromosomes for prokaryotic and eukaryotic organisms are similar in structure’ to be wrong. Therefore this discovery for which in 1993 the Nobel Prize in Physiology or Medicine was awarded is an example of a ‘Challenge’ discovery. This discovery is an example of a discovery where after the primary discovery a second breakthrough moment occurred at a later stage, when application-oriented science took off. This case study is covered in Chapter 5.

1.4.6 Validation study (Chapter 6 of this thesis)

The results of the four case studies ‘Integrase Inhibitors’, ‘Graphene’, ‘Ubiquitin’, and ‘Introns’ were used to develop and construct five algorithms. The purpose of this validation study is to validate the algorithms, which is done by comparing the outcomes of applying the algorithms to publications published in the period 1990–1994. For the validation several external datasets such as Nobel Prizes awarded, Nature’s list with the top-100 publications most cited ever, and references from review publications, patents, and social media messages were used. Publications are classified either as ‘non-breakout’ when it is not selected by one of the algorithms, or as a ‘breakout’ when it is. Based on the characteristics of the breakout publications this group is subdivided into the next four subgroups (1) ‘normal’ breakouts, (2) Breakthrough by proxy’, (3) Science-oriented breakthrough by proxy and (4) Technology-oriented breakthrough by proxy. Further details of this classification are provided in Section 6.A.9.

1.5 Information sources

1.5.1 Research publications

The bibliographic information in the WoS contains, for every publication, the list of authors, the title, publication year, the journal in which the publication was published, the referenced publications and the publications that

cite a publication and the like. At CWTS³⁵ the WoS database is enriched with information produced by CWTS's internal software. A comparable data-infrastructure for the SCOPUS-database with the same quality level is not available at CWTS at this moment; therefore the in house version of WoS-database at CWTS (CWTS-WoS) is used in this study. This CWTS-WoS database contains bibliographic information of more than 44,000,000 unique scholarly publications, published since 1980. The publications have different types, and all document types are combined into four classes (1) scholarly articles, (2) letters, (3) review articles, and (4) 'other' publications. In the CWTS-WoS database, the following information, used in this study, is available for the individual publications:

- Authors — names, addresses
- Author affiliations — names, addresses
- Science field, indicated by WoS subject categories, and also by CWTS-document clusters
- Research focus of journals³⁶
- Citation links between publications
- Publication date
- Title
- Abstract

This study focuses on the development and diffusion of scientific knowledge and, in particular, on the identification of scholarly publications having a more than average impact on the further development of a scientific field. Original research is published in publications that are classified in the WoS as document types 'article' or 'letter'. Review publications constitute another significant category of scientific publications and provide an overview of the most relevant research in a specific field of science. Therefore these publications do not contain results of original scientific research. Of the publications from the period 1990–1994, 76%³⁷ are scholarly publications, of which 91.5% are of type 'article', 6.0% of type 'letter', and 2.5% of type 'review publication'.

1.5.2 Patent publications

Patent data is used in the validation study (Chapter 6). The European Patent Office provides a database containing patent related information. This database is generally referred to as the 'PATSTAT' database³⁸, and is applied in the validation study (see Chapter 6). Twice a year a snapshot of the EPO master

³⁵CWTS - Centrum voor Wetenschap- en Technologie Studies (Centre for Science and Technology Studies), Leiden University, Leiden, the Netherlands

³⁶At CWTS a classification is assigned to journals to express the scientific orientation of the research published in the journal. Publications inherit the science orientation of the journal. The classification consists of the six classes 'Discovery science', 'Industrially relevant science', 'Science-based technological development', 'Clinically relevant science', 'Science-based clinical practice', and 'Industrial/medical development'. The methodology is explained in (Tijssen, 2010).

³⁷The 24% of 'other' publications are news items, corrections, editor notes, and such like.

³⁸The official name for PATSTAT is 'the EPO Worldwide Patent Statistical Database'

documentation database is taken to produce a new version of the PATSTAT database. The most recent version (Fall 2016) contains bibliographic data for over 90 million patent applications from more than 80 regional or national patent offices. For all patent publications the database contains information on:

- Patent applicants — names, addresses
- Inventors — names, addresses
- Various dates — date of application, of publication, of grant, and of lapse
- Area of technology — based on the fine grained patent classification systems
- Citation links to other patents
- Citation links to so called *non-patent literature*, e.g. scholarly publications
- The origin of a citation link — applicant, employee of a patent office, third party
- Title and abstract of an invention
- Information on patent families with equivalent documents, or families of technology related inventions

1.5.3 Measured quantities

Counting publications, citations, and patents

We use numbers of publications and numbers of citations at certain moments in time as data points. To study the evolution of science fields, the flow of publications is used as an approximation of the dissemination of the knowledge related to a scientific finding and of the way other researchers follow-up on this finding.

In the case of patents the focus is on a group of identical patent publications that describe the same invention, so-called patent families. In this study patent families that follow the DOCDB³⁹ definition are used to circumvent the problem of multiple counting of the same invention.

Publications of type ‘article’ and ‘letter’

Original research is published in papers that are classified in the WoS as ‘article’ or ‘letter’. Of these publications from the period 1990–1994 76%⁴⁰ are scholarly publications, as already mentioned in Section 1.5.1 on page 19 of which 91.5% is of type ‘article’, 6.0% of type ‘letter’, and 2.5% of type ‘review publication’. As the number of review papers is small in relation to the total number of publications, not including the review papers in the analysis has

³⁹Related patent publications can be grouped together in several ways. A group containing, in patent-legal sense, equivalent patent documents is called a DOCDB patent family. Patent documents in the same DOCDB-family describe the same invention.

⁴⁰The 24% of ‘other’ publications are news items, corrections, editor notes, and the like.

no significant influence on the results. In this study the classification of publications in the WoS is not questioned.

Linking patents and scholarly publications

In the PATSTAT database 15% of the total number of citations in patent publications are citations to non-patent literature⁴¹; half of these citations are references to scholarly publications. Especially in the case of technologies that are considered very promising, like Graphene, there is often at least one publication from the USPTO present in the patent family, a group of identical patent publications that describe the same invention⁴². In this study, patent families that follow the DOCDB⁴³ definition are used to circumvent the problem of multiple counting of the same invention. Of all DOCDB families in the PATSTAT database, approximately 20% contain at least one USPTO patent publication. In Figure 1.1 an example of the first page of a USPTO publication (US 9,315,437) is shown. The use of patent families solves, at least in part, the problems with references to scholarly literature in patent publications (Jaffe and de Rassenfosse, 2016).

WoS subject categories and CWTS document-clusters

WoS subject categories group together publications in 251 different science research fields. Each of these subject categories identifies a research field in science. If the subject categories change as a result of changes in science, the bibliographic information for the publications is adapted to show the actual situation.

A document classification system has been developed recently at CWTS (Waltman and van Eck, 2012) as an alternative to the WoS Subject category system. This CWTS clustering-method groups publications together on the basis of citation relations. This CWTS-system is a hierarchical system with three levels of aggregation: micro, meso, and macro. At the meso level, the level used in this study, there are 865 different 'document clusters'. Whereas the WoS-subject categories are stable over time, the CWTS document clustering reflects the dynamic nature of science as this system is based on citation relations.

⁴¹Non-patent literature consists of scholarly publications, handbooks, international standards, and everything else that is considered from a patent procedural perspective to be a reference to non-patent information. Current research at CWTS shows that roughly 50% of the references to non-patent literature are references to scholarly publications

⁴²Due to the US legislation patent publications published by the US Patent and Trademark Office (USPTO) contain extensive lists of references to all relevant - patents and non-patent-information. In patent publications provided by other patent offices usually only citations occurring in 'search' or 'examination' reports are mentioned; the citations mentioned in these reports constitute only a small portion of the citations in the patent documents.

⁴³In several ways related patent publications can be grouped together. A group containing, in patent-legal sense, equivalent patent documents is called a DOCDB patent family. Patent documents in the same DOCDB-family describe the same invention.



US009315437B2

(12) **United States Patent**
Budzik et al.

(10) **Patent No.:** **US 9,315,437 B2**
(45) **Date of Patent:** ***Apr. 19, 2016**

(54) **LOW MOLECULAR WEIGHT CATIONIC LIPIDS FOR OLIGONUCLEOTIDE DELIVERY**

(71) Applicant: **SIRNA THERAPEUTICS, INC.**,
Cambridge, MA (US)

(72) Inventors: **Brian W. Budzik**, Perkiomenville, PA (US); **Steven L. Colletti**, Princeton Junction, NJ (US); **Jennifer R. Davis**, Richboro, PA (US); **Ivory D. Hills**, Deerfield, MA (US); **Daria Danile Seffried**, Lansdale, PA (US); **Matthew G. Stanton**, Marlton, NJ (US)

(73) Assignee: **Sirna Therapeutics, Inc.**, Cambridge, MA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.
This patent is subject to a terminal disclaimer.

(21) Appl. No.: **14/711,246**

(22) Filed: **May 13, 2015**

(65) **Prior Publication Data**
US 2015/0315112 A1 Nov. 5, 2015

Related U.S. Application Data

(63) Continuation of application No. 14/265,029, filed on Apr. 29, 2015, now Pat. No. 8,062,021, which is a continuation of application No. 13/701,636, filed as application No. PCT/US2011/038490 on May 31, 2011, now Pat. No. 8,748,667.

(60) Provisional application No. 61/351,373, filed on Jun. 4, 2010, provisional application No. 61/382,067, filed on Sep. 13, 2010.

(51) **Int. Cl.**
C07C 43/15 (2006.01)
C07D 295/084 (2006.01)
C07C 217/44 (2006.01)
C07D 317/28 (2006.01)
C07D 267/08 (2006.01)
C07D 205/04 (2006.01)
C07C 217/46 (2006.01)
A61K 9/127 (2006.01)

(52) **U.S. Cl.**
CPC **C07C 43/15** (2013.01); **A61K 9/1272** (2013.01); **A61K 47/18** (2013.01); **C07C 217/44** (2013.01); **C07C 217/46** (2013.01); **C07D 205/04** (2013.01); **C07D 267/08** (2013.01); **C07D 295/084** (2013.01); **C07D 317/28** (2013.01); **C12N 15/87** (2013.01)

(58) **Field of Classification Search**
None
See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

6,060,455 A	5/2000	Fernholz et al.
8,748,667 B2	6/2014	Budzik et al.
9,062,021 B2 *	6/2015	Budzik et al.
2009/0285881 A1	11/2009	Dunde et al.
2010/0055168 A1	3/2010	Dunde et al.
2010/0063135 A1	3/2010	Dunde et al.
2010/0099738 A1	4/2010	Hansen et al.

FOREIGN PATENT DOCUMENTS

WO	2009129385	10/2009
----	------------	---------

OTHER PUBLICATIONS

Database CAPLUS on STN, Acc. No. 1999-65895, Bissel et al., Langmuir (1999), 15(5), p. 1791-1795.
Semple et al. "Rational design of cationic lipids for siRNA delivery", Nat. Biotechnol. 28(2): 172-6 p. 713, Fig. 3C", Feb. 2010.

* cited by examiner

Cited patent publications

* cited by examiner **Cited scholarly publications**

Primary Examiner — Brian J Davis
(74) *Attorney, Agent, or Firm* — Nixon Peabody LLP

(57) **ABSTRACT**
The instant invention provides for novel cationic lipids that can be used in combination with other lipid components such as cholesterol and PEG-lipids to form lipid nanoparticles with oligonucleotides. It is an object of the instant invention to provide a cationic lipid scaffold that demonstrates enhanced efficiency along with lower liver toxicity as a result of lower lipid levels in the liver. The present invention employs low molecular weight cationic lipids with one short lipid chain to enhance the efficiency and tolerability of in vivo delivery of siRNA.

6 Claims, 2 Drawing Sheets

Figure 1.1: Front page of United States patent 9,315,437

Chapter 2

Case study #1: Integrase Inhibitors

The content of this chapter is equivalent to the manuscript published as:

Winnink, J.J. and Tijssen, R.J. (2014). R&D dynamics and scientific breakthroughs in HIV/aids drugs development: the case of integrase inhibitors. *Scientometrics*, 101(1):1-16. (DOI:10.1007/s11192-014-1330-7).

The only differences between the published article and the text in this chapter are of typographic nature to assure that the layout is in line with the typographic design of this PhD thesis.

Abstract

Transformations and applications of scientific knowledge into new technologies are usually complex interactive processes. Is it possible to detect from bibliographic information alone, structural alterations and significant events within these processes that may indicate breakthrough discoveries? In this empirical study we focus on R&D processes leading to HIV/AIDS medicines called Integrase Inhibitors. Where scientific progress and discoveries are reflected in research papers, patents signify inventions and technological achievements. Our temporal analysis of distinctive events in this R&D area, tracing trends within both bibliographic information sources, is driven by three bibliometric indicators: (1) contributions of 'bridging researchers' who are also inventors, (2) 'key papers' that subject experts in the field considered milestones in the research process, and (3) multidisciplinary impact of those papers. The main results indicate that a combination of key papers, bridging researchers and multidisciplinary impact might help track potential 'Charge type' breakthrough developments.

2.1 Introduction

This paper reports on the development of a statistical method for data-mining bibliographic information in order to describe and analyse ‘distinctive’ events or transformative developmental patterns within R&D processes. The main goal is to assess whether or not such events or transformations represent ‘breakthroughs’, i.e. the introduction of remarkably new knowledge or radically innovative technologies. There is growing academic literature discussing the characteristics of such breakthroughs, e.g. (Ahuja and Lampert, 2001; Chandy and Tellis, 2000; Chen et al., 2009; Henderson and Clark, 1990; Hollingsworth, 2008; Podolny and Stuart, 1995; Rosenkopf and Nerkar, 2001; Trajtenberg, 1990a,b; Tushman and Anderson, 1986). In this empirical study we focus our attention on scientific discoveries that are the cause of, or significantly contribute to, subsequent technological developments.

Scientists, researchers and scholars are often able, retrospectively, to identify documented evidence of such discoveries or inventions that were of crucial importance in these R&D processes. Such a document will be referred to as a ‘*key publication*’. These particular publications provide empirical information to help trace and identify transformative stages within R&D processes where a breakthrough occurred, and help describe characteristic features and structural determinants of processes that played a crucial role. In this study we examine the development of HIV/AIDS medicines, more specifically the research area of ‘Integrase Inhibitors’. Our main research question is: ‘Which indicator or indicators, based on bibliographic data of scholarly publications and patent publications, could be applied in early stage identification of potential scientific breakthroughs?’ Section 2.2 ‘Analytical framework’ first briefly introduces previous work on identifying breakthroughs, discusses the concept of ‘bridging researchers’, provides a short overview of the development of HIV/AIDS medicines, and argues the decision to focus in this study on the development of Integrase Inhibitor medicines. We further describe our bibliographic information sources, and discuss how longitudinal developments were analysed. The main results in Section 2.3 ‘Results’ highlight highly cited key publications in this research area, and the multidisciplinary impact against the backdrop of the most prominent bridging researchers. Section 2.4 ‘Discussion and concluding remarks’ summarizes and concludes with general observations.

2.2 Analytical framework

2.2.1 Defining and identifying breakthroughs

There are many conceptual and methodological challenges involved in identifying scientific breakthroughs, and pinning down the moment within the R&D process when they occurred. Various detection methods and analytical tools already have been tried and tested to solve this problem — with limited success (Arbesman, 2010; Bettencourt et al., 2008; Breiner et al., 1994; Julius

et al., 1977; Martin, 1995). More recently, Chen (2012) takes an ‘informational’ perspective by focussing on the question ‘What kinds of information may serve as early signs of potentially valuable ideas?’; Small and Klavant (2011) combine co-citation analysis and the context of citations; Ponomarev et al. (2012) provides an illustrative example of this data-informed and evidence-based approach by means of a two-step citation impact methodology applied to research publications.⁴⁴

Our informational study also operationalizes (potential) breakthroughs in terms of bibliographic data, but from a broader range, which includes research publications and patents. In our longitudinal study time is of the essence. We define an event time-line based on the publication date of the research publication and the oldest priority date of patents.⁴⁵ Within this framework we characterize key events by the emergence of one or more key papers that were produced by bridging researchers. These papers are either very highly cited, and/or exhibit a broad multidisciplinary citation impact, within 12 months after publication. Our underlying conceptual framework is based on Hollingsworth’s description of a breakthrough ‘A major breakthrough or discovery is a finding or process, often preceded by numerous small advances, which leads to a new way of thinking about a problem’ (Hollingsworth, 2008), and on Koshland’s Cha-Cha-Cha typology of scientific discoveries (Koshland, 2007).⁴⁶ Koshland distinguishes three types of discoveries: ‘**Challenge**’ discoveries are a response to an accumulation of facts or concepts that remained unexplained by, or were incongruous with, scientific theories of the time, ‘paradigm shifts’ (Kuhn, 1962); ‘**Chance**’ discoveries suddenly appear and are often referred to as serendipitous, (van Andel, 1994) denotes discoveries of this type ‘unsought findings’; ‘**Charge**’ discoveries solving known problems within an existing theoretical framework can be considered, in Kuhnian sense, evolutionary ‘normal’ science. Making a distinction between charge discoveries that do not involve a paradigm shift and challenge and chance discoveries that do is relevant as paradigm-changing discoveries have notoriously limited early impact (Redner, 2005), and are therefore hard to detect on the basis of early citation patterns (Wang et al., 2013). The Integrase Inhibitors breakthrough is a *Charge* discovery; the biochemical mechanism was already identified and no new theoretical understanding was needed — finding the right chemical compound to do the job was the main problem.

2.2.2 Bridging researchers

In addition to citation-based linkages between key publications, one can also connect documents through authorship. The same people can act as (co)

⁴⁴Tracking the citation patterns over a period of 6–24 months after publication, and using this information to predict if a publication exceeds a citation threshold level within 5 years after publication.

⁴⁵Oldest priority data is the date of the earliest filing of a patent application for this invention.

⁴⁶Ciechanover (2009) applies Koshland’s criteria to motivate the classification of the discovery of ‘ubiquitin proteolytic system’ as a ‘**Challenge**’ rather than a ‘**Chance**’.

author of a research publication and as (co-)inventor of a patent. These author/inventors constitute an explicit link ('bridge') between knowledge creation and knowledge application. We argue that these 'bridging researchers' may play a pivotal role in overcoming barriers between basic research, applied research, and technological development in the development of radically new medicines. They are referred to as 'core scientists' (Furukawa and Goto, 2006) or 'Pasteur scientists' (Stokes, 1997) owing to their ability to recognize and reconcile different stages in the R&D chain. They have been subject of several studies of knowledge transfer mechanisms, where these individuals are considered to be an important thrust in moving scientific breakthroughs towards science-based innovations (Balconi et al., 2004; Lissoni, 2010; Meyer, 2006; Noyons et al., 1994; Tijssen, 2002; Zucker and Darby, 1996).

2.2.3 Development of new HIV/IDS medicines

Usually many years may pass before scientific discoveries may find their way into technological applications (Isenson, 1969; Jewkes et al., 1958). In medicine, it often takes 10-14 years from 'upstream' disease characterization to a 'downstream' drug approved for human usage (IOM, 2009). The costs involved in developing a new medicine from start to approval are estimated (FDA, 2004) to be in the range from US\$ 0.8 to 1.7 billion. The public sector is key contributor in the process of drug development in the sense that universities or government laboratories conducts 'enabling research' (Cockburn and Henderson, 1998). In the 1980s a novel method for drug design known as 'Rational Drug Design' (RDD)⁴⁷ replaced then existing trial-and-error 'harvesting' methods for finding new 'lead compounds'⁴⁸ that could act as a basis for a new medicine. RDD's encouraged the development of various computational methods using structural information suggesting novel molecular structures. Each of these structures may either prove to be useful as a 'lead compound' or act as a stimulus to the creativity of designers (Wang et al., 2000). Using RDD the route from scientific knowledge to applicable HIV/AIDS medicines was traversed in a relatively short period of time and the development of these medicines has been very dynamic since the very start in the early 1980s.

HIV/AIDS medicines are currently classified into five⁴⁹ distinct classes based on the bio-molecular mechanisms used to attack the pathogen. In his review paper De Clercq (2009a) identifies a series of key publications describing developments considered being pivotal for the developments in the classes of HIV/AIDS medicines. Owing to the controlled and sequential nature of RDD-based development processes of new medicines, we should be able to identify and study scientific breakthrough 'transformative' events. We focus on R&D for medicine in the class of Integrase Inhibitors and specific on a

⁴⁷Also known as 'computer-aided drug design', or 'structure-based drug design',

⁴⁸Lead compounds are, if proven to be an effective drug, the beginning of a group of chemical closely related medicines

⁴⁹A sixth class is formed by the so called cocktail drugs; combinations of two or more drugs from the other five classes.

2.2 Analytical framework

medicine with the generic name Raltegravir, which is also known by its brand name *Isentress*[®]. It was the only approved drug of this class at the time of this study.

2.2.4 Focus on Integrase Inhibitors

The rationale to study the development of Integrase Inhibitors medicine is based on the following considerations: (1) the key paper that is the basis of the *Integrase Inhibitors* medicines (Hazuda et al., 2000) stands out, as it is the only ‘citation classic’ among the HIV/AIDS key papers (see Section 2.3), (2) at the time of this study *Isentress*[®] was the only approved drug of this class, and (3) the relative short period between the publication of the key paper on January 28, 2000 and the publication of the granted patent on January 30, 2007. Especially (2) and (3) restricted the number of scholarly and patent publications in the analysis, and therefore reducing the complexity of the study.

2.2.5 Bibliographic data

We selected bibliographical data for relevant scholarly publications from the period 1980–2010 from our Thomson Reuters/CWTS Web of Science database (TR/CWTS WoS). This dataset contains bibliographical information on 34,665 scholarly publications consisting of articles and letters related to the disease HIV/AIDS and to the development of HIV/AIDS medicine. We restricted the dataset to a number of subject categories⁵⁰ in order to exclude references to non-pharmaceutical or non-medical subject areas.

The patent data were extracted from the April 2012 release of the EPO PATSTAT database⁵¹ covering the period 1980–2008 consisting of 16,379 distinct so called DOCDB patent families⁵². Each patent family was counted only once. We constructed the set of patent publications by applying patent classification codes related to HIV/AIDS, which was supplemented by with documents containing the terms ‘HIV’, or ‘AIDS’. We imposed the constraint that the patent publications were also classified as ‘medical preparation’ (i.e. a medicine). In this set of documents we identified all patents belonging to the *Integrase Inhibitors* class.

Patent publications frequently contain references to non-patent publications. Part of those non-patent references (NPR’s) consists of references to scholarly publications, which reflect direct or indirect links between the cited scientific knowledge and the citing patented technology, see e.g.(Grupp, 1992;

⁵⁰Immunology, biochemistry & molecular biology, virology; cell biology, microbiology, infectious diseases, pharmacology & pharmacy, biotechnology & applied microbiology, medicinal chemistry, infectious diseases, and medicine

⁵¹The PATSTAT database is the database also known as the *EPO Worldwide Patent Statistical Database*.

⁵²A DOCDB patent family is a collection of closely related equivalent patent publications describing the same invention. 2008 was used as cut-off year because due to procedural regulations more recent patent data were not yet fully available.

Hicks, 2000; Narin and Olivastro, 1998; Tijssen, 2002). These NPRs were used to study the temporal dynamics of science being utilized and transformed into patented applications and technological breakthroughs.

In order to track micro-level developments over time we need to synchronize the time line between bibliographic evidence. For patents we adopted ‘the oldest priority date’, the available date that is closest to the moment of the invention. For scholarly publications we selected the date a manuscript was received by the publisher, acknowledging the fact that this date is a crude proxy for the moment the breakthrough idea actually came about.⁵³

Our case study focuses on one key publication that played a pivotal role in this development: the ‘Hazuda 2000’ paper that published the solution to the then remaining scientific problems (Hazuda et al., 2000). We identified all patent publications that contain a literature reference to this publication, which enabled us to compile a set of patent documents directly influenced by this publication. Starting with the patent publication on *Isentress*[®] (US patent 7169780) we created a backward citing-cited network and collected the publication years for the referenced publications. In the Web of Science, we selected publications published from 1999 onwards, which allowed us to include information prior to the hard-copy publication of ‘Hazuda 2000’. Drawing on these sources, all inventors, authors, patent applicants, companies or research institutions active in the *Isentress*[®] R&D area were identified and incorporated in a longitudinal network structure.

2.3 Results

General

HIV was identified 1983 as the retrovirus causing the disease AIDS. The first research papers related to the development of HIV/AIDS medicine were published in 1981. Following the initial HIV/AIDS medicines five new classes of medicines based on new insights in the bio-molecular workings were developed. For each of these classes De Clercq’s authoritative, Table 2.3 contains a short biography, review highlights the research publications that mark pivotal developments (De Clercq, 2009b). Table 2.1 lists all seven of De Clercq’s key publications which we used as a frame of reference to assess (Hazuda et al., 2000) as a possible breakthrough. Three publications (Baba et al., 1989; Miyasaka et al., 1989; Pauwels et al., 1990) constitute the start of the non-nucleoside reverse transcriptase inhibitors.

Figure 2.1 shows the number of citations received by the key papers and an (weighted) average HIV/AIDS publication received the first 10 years after

⁵³This manuscript submission date is however not provided in the databases. For individual publications the submission date of the manuscript can usually be found on the ‘image’ of the publication, but for analysis of larger datasets the use of this date is very laborious and time-consuming, and we therefore had to use the ‘publication date’ provided by the Web of Science. Further research has to be done in this area to estimate the error introduced; (Luwel and van Wijk, 2012) provides some insights into the nature and extent of this problem.

2.3 Results

Table 2.1: General description used by De Clercq (2009a) to mark pivotal developments supplemented by the number of citations received by the key papers

Publication	Description of the pivotal development (De Clercq, 2009a)	Citations received		
		From scholarly publications		From patent publications
		During first 12 months	Total (ultimo 2010)	Total (ultimo 2008)
Baba et al. (1989) ^a	Discovery of chemical substance later to be known as HEPT; start of so called non-nucleoside reverse transcriptase inhibitors (NNRTIs)	3	265	0
Miyasaka et al. (1989)	Discovery of chemical substance later to be known as HEPT; start of so called non-nucleoside reverse transcriptase inhibitors (NNRTIs)	7	201	8
Pauwels et al. (1990)	Discovery of chemical substance later known as TIBO; start of so called non-nucleoside reverse transcriptase inhibitors (NNRTIs)	29	571	26
Roberts et al. (1990)	Discovery of Saquinavir, first of the proteinase inhibitors (PIs)	32	570	88
Wild et al. (1993)	Discovery of Enfuvirtide, first drug in the class fusion inhibitors (FIs)	3	115	12
Hazuda et al. (2000)	Scientific base for integrase inhibitors (INIs)	16	585	9
Esté and Telenti (2007) ^b	Discovery of mechanism on which co-receptor inhibitors (CRIs) are based	22	114	0

^a Baba et al. (1989); Miyasaka et al. (1989) both concern the same finding (HEPT)

^b Esté and Telenti (2007) is actually a review article, not specifically dealing with a certain drug

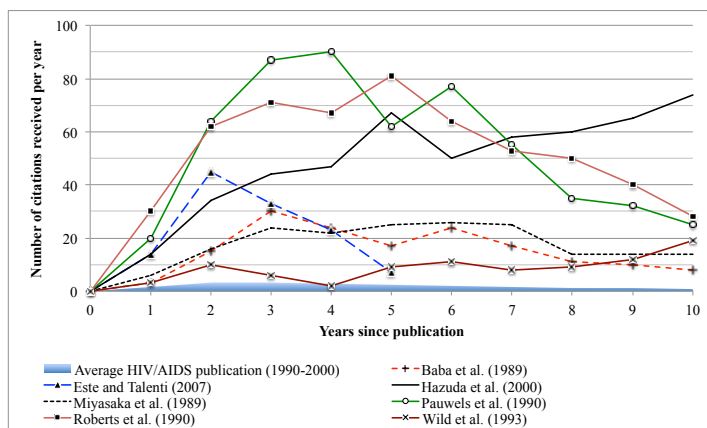


Figure 2.1: Number of citations received per year by the key papers and an average HIV/AIDS publication during the first 10 years after publication

publication as well as the (weighted) average of all HIV/AIDS publications. To compute the weighted average we only included documents that were cited at least once during the period. Each of these seven key papers receives significantly more citations than average. 'Hazuda 2000' is unique in its continued high-volume citation impact, which makes it the only 'citation classic' in its field.

Figure 2.2 exhibits the general trend in the volume of scholarly publications and patent applications related to the class of 'Integrase Inhibitors' medicines. Scholarly publications emerged in 1991. Patent publications took off in 1997.⁵⁴ In other words, it took 6 years to utilize and transform scientific knowledge into a patentable technology; even the more efficient 'RDD' approach to develop new drugs is still a complicated and time-consuming trajectory. The graph also shows a steep increase in the number of publications in 2001, after 'Hazuda 2000' went public. A second steep rise occurred in 2005 following the publication of two other highly cited publications by Hazuda: (Hazuda et al., 2004a,b) receiving 12 respectively 14 citations during the first 12 months of their existence. These two belong also to the top 5 % most cited publications in the research field of Integrase Inhibitors.

Table 2.2 shows, for three different citation windows, the relative position of each key paper within the distribution of all HIV/AIDS papers published in the same year. We argue that if a research paper truly represents a scientific breakthrough, is of great relevance for further research, and it is recognized as such by peers working in the same research field, one may expect to see a relatively quick response from peers in the field and steep

⁵⁴Although 'Hazuda 2000' is seen as an important publication it is just a link in the R&D process that started before this publication, and already produced some patents. The publication of this paper might have been intentionally delayed in order not to infringe with ongoing patent applications. The bibliographic data is insufficient to solve this issue.

2.3 Results

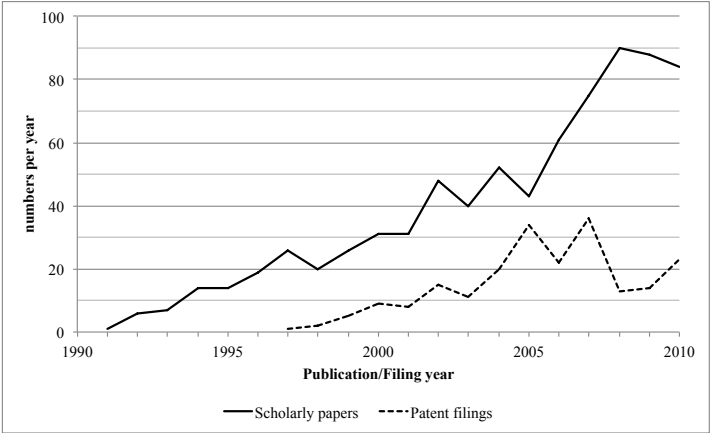


Figure 2.2: Cumulative share of publications as function of the number of citations in the first 12 months after publication

Table 2.2: Relative position of the key papers within the citation distribution of all HIV/AIDS papers published in the same year*

<i>Publication</i>	Citation window		
	1 year	3 years	10 years
Baba et al. (1989)	top 39%	top 7%	top 5%
Miyasaka et al. (1989)	top 25%	top 7%	top 4%
Pauwels et al. (1990)	top 5%	top 1%	top 1%
Roberts et al. (1990)	top 1%	top 1%	top 1%
Wild et al. (1993)	top 33%	top 17%	top 8%
Hazuda et al. (2000)	top 4%	top 2%	top 1%
Este and Talenti (2007)	top 4%	top 2%	not available

* Only citing publications of the document types 'article' and 'letter' are included

rise in citations to the paper. For early identification of such 'hot papers' we restrict our time window to year 1 after publication. So how 'hot' was 'Hazuda 2000' Figure 2.3 shows the cumulative share of the citing documents as function of the number of citations received during the first 12 months for all HIV/AIDS publications and for all 'Integrase Inhibitor' publications in our sample, where the latter is split into citations prior to 'Hazuda 2000' (indicated by '< 2001'), and citations afterwards until the follow-up publications of Hazuda et al. (2004a,b). While 65 % of the Integrase Inhibitors publications received less than 4 citations during that 1st year, 'Hazuda 2000' managed to collect 16 citations, immediately making it one of the most cited in the field. Based on the number of citations received during those first 12 months, four of the seven key papers belong to the top 5 % most highly cited HIV/AIDS publications worldwide.

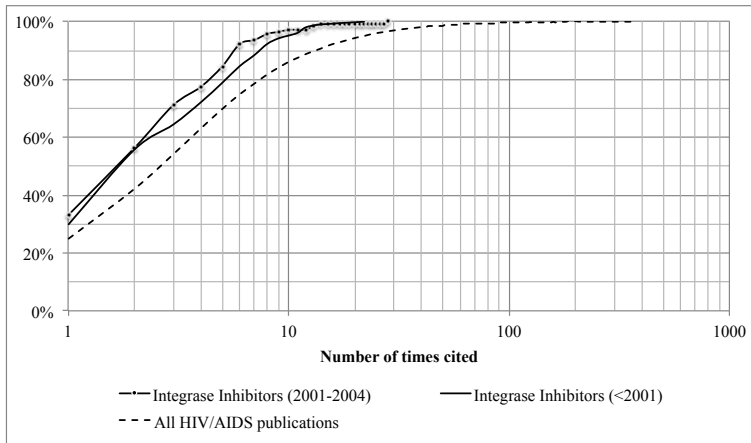


Figure 2.3: Trend of total number of distinct CWTs subject categories citations originate from

2.3.1 Multidisciplinarity of citations

A second assumption relates to the breadth of a publication's citation impact. To help identify a breakthrough we looked at the evolution of the multidisciplinarity⁵⁵ We looked at the multidisciplinarity of the citing publications, where multidisciplinarity is defined as the number of different Subject Categories⁵⁶. Figure 2.5 displays the number of additional Subject Categories for each subsequent year. The multidisciplinary citation profile (Figure 2.4) of (Hazuda et al., 2000) is comparable to (Pauwels et al., 1990) and (Roberts et al., 1990). The latter two actually marshal the largest disciplinary spectrum, a result of three waves of expansion in years 3, 6 and 10. What then makes 'Hazuda 2000' so distinctive in terms of the sheer volume of citations? And why its continued high-level citation impact?

2.3.2 Bridging researchers

Of all authors publishing in the field of Integrase Inhibitors only 5.4% are such bridging researchers. Within the patent publications we found that 30.6% of the inventors belong to this select group, which clearly indicates that scientific research and drug development are closely linked in this field. Our analysis of the patents also reveals that all first authors of the seven key papers (Table 2.1) are bridging researchers. De Clercq (2013) commented 'Daria

⁵⁵Porter and Rafols (2009) and Ponomarev et al. (2012) use the related term 'interdisciplinarity' to classify research publications that cover several scientific disciplines.

⁵⁶CWTs uses 251 distinct subject categories based on the 'standard' subject categories of used by Thomson-Reuters in the Web of Science database; 60 of these CWTs subject categories are related to medical sciences and life sciences; the themes range from fundamental biomedical research to clinical medicine. It is expected that citations start to come a increasing and more 'application' oriented range of subject categories as research progresses

2.3 Results

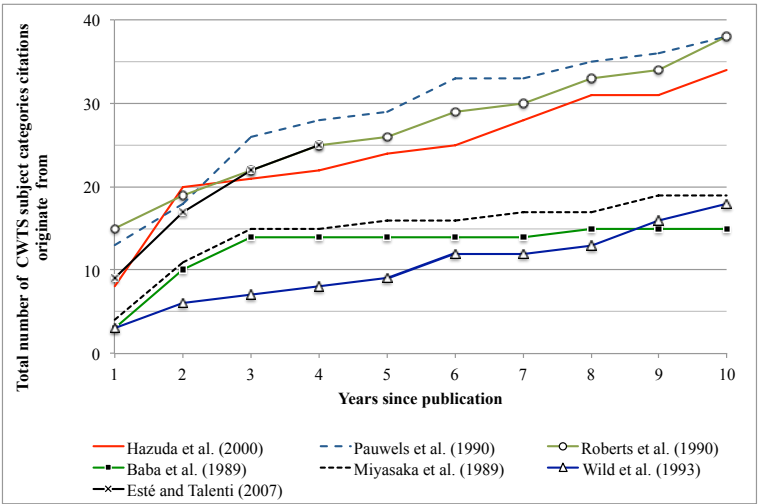


Figure 2.4: Numbers of ‘new’ subject categories citations originate from added in a year

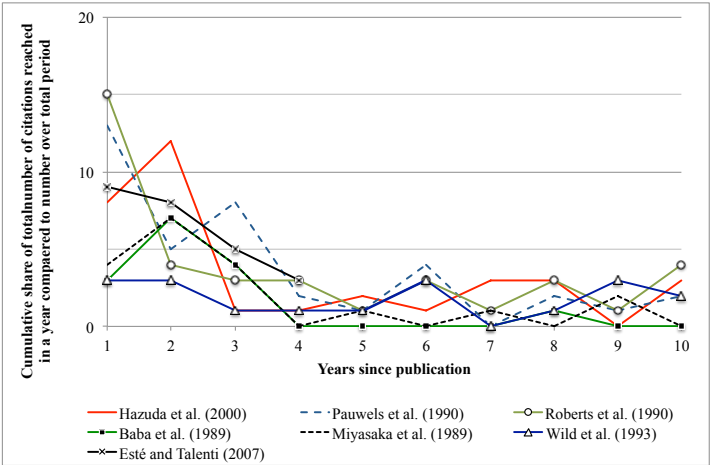


Figure 2.5: Trend of scholarly and patent publications in the field of ‘Integrase Inhibitors’

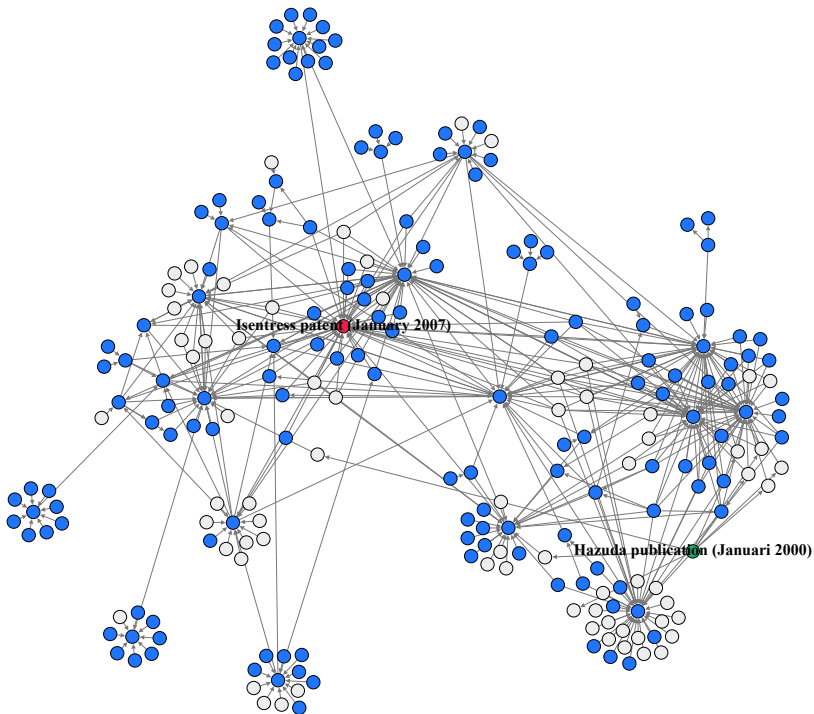


Figure 2.6: Network of patents and scholarly publications connection the Hazuda paper and the Isentress® patent. Blue (opaque discs) discs represent patent publications; white discs represent scholarly publications

2.3 Results

Hazuda is a unique example of a bridging researcher, as she is at the origin of the concept (strand transfer integrase inhibitors) as well as the drug [Raltegravir (*Isentress*[®])].'

If this paper is indeed such an influential contributor, it should occupy a core node in knowledge creation and utilization networks. Does it? These networks are reflected in citation linkages within and between sets of publications and patents. We compiled a set of patent publications with a citation link to this particular patent. No less than 18 companies and organisations applied for patents that are directly linked to this paper by explicit citation relationships. One of these companies is *Merck & Co* where Hazuda's research was done; one of its subsidiaries owns the *Isentress*[®] patent. This set of citing patents contains 135 distinct inventors. On the scientific side we identified 1,878 (co-) authors citing 'Hazuda 2000'. All in all, both sources collectively assemble sufficient information to construct the citing-cited network of scholarly publications and patent publications depicted in Figure 2.6. Taking the *Isentress*[®] patent as a starting point, and following backward citation links to either other patent publications or scholarly publications, we collected cited documents. The search process terminated when a scholarly publication was found or when either the publication date of a publication or the filing date of a patent publication reached the year 1999. Interestingly we found no direct citation links between 'Hazuda 2000' and the *Isentress*[®] patent. However they do connect in seven cases via intermediate citing-cited links. Six of these intermediates are scholarly publications, including Hazuda et al. (2004a,b), and one is a patent publication. This loose coupling between breakthrough science and patented medicine reflects stepwise processes where 'Hazuda 2000' reports on an initial 'enabling discovery', only one of several building blocks needed to create a new medicine, where Hazuda's follow-up papers in 2004 are more application oriented.

We identified 21 persons as bridging researchers. In Figure 2.7 we present these bridging researchers and the publications related to the development of *Isentress*[®] they (co) authored. This figure shows the science-technology border in case of the *Isentress*[®] development based on the citing-cited relations between the publications and links based on (co) authorships relationships. Two clusters are visible; the larger cluster is centred around the 'Hazuda paper'. The three publications by Hazuda (Hazuda et al., 2004a, 2000, 2004b) in this cluster received respectively 16, 12, and 14 citations during the first twelve months after publication.

The other publications in this cluster received three or fewer citations in the same period. This cluster presents the scientific R&D that played a key role in the development of *Isentress*[®]. Two papers (Hazuda et al., 2004a,b) have seven bridging researchers in common and are linked by a citing-cited relation to the Hazuda paper. Hazuda is (co) author of these three highly cited publications and plays therefore a central role. The papers in the second cluster received only a small number of citations.

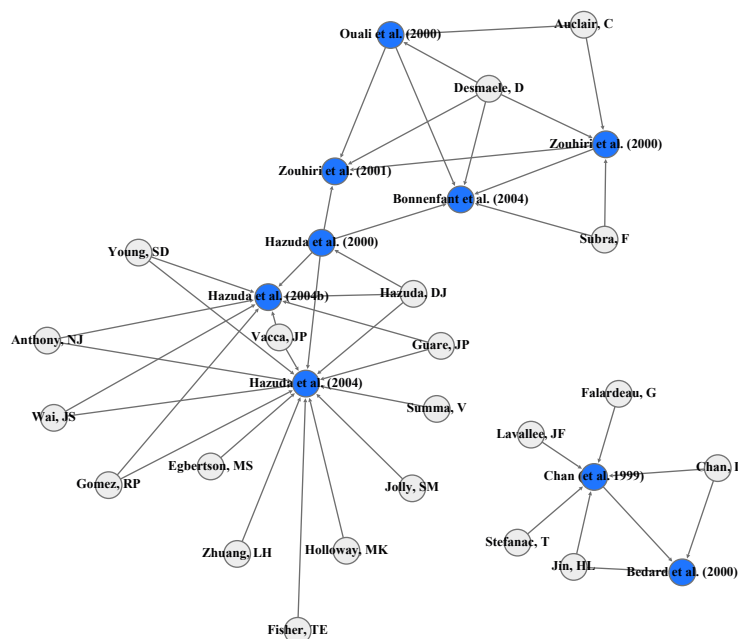


Figure 2.7: Networks of bridging researchers and their publications related to *Isentress*®. Blue (opaque discs) discs represent scholarly publications; white discs represent (co) authors

2.4 Discussion and concluding remarks

This paper develops and suggests a novel analytical framework for early-stage detection of potential scientific breakthroughs. Its applicability hinges critically on several methodological assumptions and technical features related to the empirical data we used and the quality of the databases. In this pilot study we applied several constraints: (1) we restricted our study to only one of the six classes of HIV/AIDS drugs; (2) the focus is on R&D processes cumulating in an approved medicine; (3) a relative short period of time is used for tracking citations; (4) all scholarly and patent publications, linked by citing-cited relations from the initial 'Hazuda 2000' paper to the final *Isentress*® patent, are included; (5) authors and bibliographical links are included insofar they could be identified in the databases.

Using extensive volumes of bibliographic information to identify key research publications and associated 'bridging' researchers, we analysed the R&D dynamics for 'Integrase Inhibitors' and identified what seem to be breakthrough stages in the R&D process. The 'Hazuda 2000' publication is of particular importance according to De Clercq (2009b), and the numbers of citations to this paper in the international scientific literature substantiate this claim.

'Hazuda 2000' is receiving, on average, each year 40 citations from scientific research publications and 10 from review articles; of the key papers in the HIV/ AIDS field it is the only 'citation classic', this reflects the fact that the publication serves as reference for new developments, and helps in 'new ways of thinking about a problem' (Hollingworth's definition). The bibliographic data we used indicates that a cascade of R&D activities was made possible as a follow-up of the Hazuda's achievements. Our exploratory study reveals that Hazuda is also one of the researchers that bridge the domains of science and technology. Hazuda is one of several researchers in the case of HIV/AIDS medicines who all made pivotal discoveries as mentioned in the seven key papers in the several research subfields.

In this case study the 'Hazuda 2000' publication is used as starting point to search for alterations in bibliographic data related to breakthroughs in the development of a specific type of HIV/AIDS medicines. According to an expert in this field of research the Hazuda paper describes a breakthrough in the development of medicines in the Integrase Inhibitors group. The main methodological conclusions of our experimental approach are that, under the right conditions, potential scientific breakthroughs in areas of science-based technology might be found by searching bibliographic data sources for highly cited 'key' publications which are (co-) authored by bridging researchers and that show an increasing multidisciplinary impact. These critical conditions depend on: (1) the ability to identify citation links between scholarly publications and patent publications; (2) the maturity of the research area in order to delineate document sets; (3) the coverage of the publications of the research field in the bibliographic databases. The method is therefore — by default — restricted by the information contained in the bibliographic databases we used; we therefore cannot rule out the possibility that we overlooked important features and contributions in R&D processes that are not covered by the databases.

Acknowledging these cautionary remarks, the results of this exploratory case study suggest that indicator-based methodologies are able to, retrospectively, find potential breakthrough research papers relatively soon after their publication in the open scientific literature. The extent to which these general observations and conclusions can be extended to other R&D areas and processes (other than science-driven drug development) remains an open question for now. Further research is needed to apply this approach across wider range of R&D domains, to assess the probability of identifying potential scientific breakthroughs within 1 to 2 years after publication. (1) the ability to identify citation links between scholarly publications and patent publications; (2) the maturity of the research area in order to delineate document sets; (3) the coverage of the publications of the research field in the bibliographical databases. The method is restricted by the information contained in the databases we use and we cannot rule out that our method misses an important discovery because data is not covered by the databases. The extent to which these conclusions can be generalised to other areas of scientific enquiry, R&D processes and technologies other than drug development, remains

an open question for now.

2.5 Acknowledgement

The authors kindly thank Prof. (em.) Erik de Clercq for commenting on a concept version of this manuscript.

Table 2.3: Short biography of professor Erik De Clercq

Professor Erik De Clercq

European Inventor of the Year 2008 in the category "Lifetime achievement"

[] De Clercq's landmark contributions to biomedical sciences include the development of new antiviral agents for the treatment of viral infections ranging from different forms of herpes to hepatitis B and HIV. In fact, De Clercq and his team are single-handedly responsible for developing some of the original drugs against HIV. They also inspired a switch from monotherapy (treatment with one kind of drug) to the use of up to four agents at a time - the "drug cocktail" that is now the standard treatment approach for HIV. He is best known for opening up the field of so-called "nucleotide analogues", which spawned the first-ever series of broad-spectrum antiviral drugs.

[] Since 1972, he has led the Laboratory for Virology at Leuven University's Rega Institute for Medical Research. De Clercq has published over 2100 papers in peer-reviewed scientific journals and has given more than 530 lectures at international congresses, conferences and symposia on various aspects of antiviral chemotherapy. An active member of countless boards and advisory committees, including the World Health Organization (WHO) Expert Advisory Panel on Virus Diseases, De Clercq is addressing the pressing issues of our times, including the threat of an H5N1 "avian flu" pandemic.

Source:

<http://www.epo.org/learning-events/european-inventor/finalists/2008/de-clercq.html>

Chapter 3

Case study #2: Graphene

The content of this chapter is equivalent to the manuscript published as:

Winnink, J.J. and Tijssen, Robert J.W. (2015). Early stage identification of breakthroughs at the interface of science and technology: lessons drawn from a landmark publication. *Scientometrics*, 102(1):113-134. (DOI:10.1007/s11192-014-1451-z).

The only differences between the published article and the text in this chapter are of typographic nature to assure that the layout is in line with the typographic design of this PhD thesis. The information presented in section 'Appendix 3.A Supplementary information — co-occurrence term maps' on page 64 is also available on the website of the publisher in a separate file accompanying the journal article.

Abstract

Certain scholarly publications or patent publications may signal breakthroughs in basic scientific research or radical new technological developments. Are there bibliographical indicators that enable an analysis of R&D dynamics to help identify these 'local revolutions' in science and technology? The focus of this paper is on early stage identification of potential breakthroughs in science that may evolve into new technology. We analyse bibliographic information for a typical example of such a breakthrough to pinpoint information that has the potential to be used as bibliographic indicator. The typical example used is the landmark research paper by Novoselov et al. (2004) concerning graphene. After an initial accumulation of theoretical knowledge about graphene over a period of 50 years this publication of the discovery of a method to produce graphene had an immediate and significant impact on the R&D community; it provides a link between theory, experimental verification, and new technological applications. The publication of this landmark discovery marks a sharp rise in the number of scholarly publications, and not much later an increase in the number of filings for related patent applications. Noticeable within two years after publication

is an above average influx of researchers and of organisations. Changes in the structure of co-citation term maps point to renewed interest from theoretical physicists. The analysis uncovered criteria that can help in identifying at early stage potential breakthroughs that link science and technology.

3.1 Introduction

Martin (1995) states “... Confronted with increasing global economic competition, policy-makers and scientists are grappling with the problem of how to select the most promising research areas and emerging technologies on which to target resources and, hence, derive the greatest benefits ...” Identifying breakthroughs in science at early stage could aid in solving this selection issue, and help in timely prioritizing R&D resources by for instance policy-makers, research organisations, or companies. The focus of our research is on finding indicators in bibliographic data that would enable us to conclude at early stage if a discovery is to be considered a potential breakthrough discovery at the interface⁵⁷ of science and technology. We study well-known cases generally recognized by experts as breakthrough discoveries in science. In this study the focus is on the publication (Novoselov et al., 2004) for which the Nobel Prize for Physics 2010 was awarded to Konstantin S. Novoselov and Andre K. Geim for “succeeding in producing, isolating, identifying and characterizing graphene” (Nobel Prize Committee, 2010). We will refer to the paper in which this discovery was published as the ‘Novoselov paper’. Graphene⁵⁸ seems to be an appropriate case in point to study because this paper acted as a bridge between basic science and applied science, and therefore marks the beginning of technological applications. Prior to this seminal paper graphene was not available as an independent material and therefore it could not be studied in experiments or used as the basis for new technological applications.

The properties of sheets of multi-atom thick layers of graphite had already been studied for a considerable period, for instance Wallace (1947) and Boehm et al. (1962). These theoretical studies focused on the physical aspects of graphite layers (sheets). Over time, the theoretical studies uncovered several properties resulting in the expectation that ‘freestanding graphene’ would be a material with a high potential for novel and unforeseen technological

⁵⁷Interface = ‘a point where two systems, subjects, organizations, etc. meet and interact’ (Oxford English Dictionary)

⁵⁸Graphene is a material that in its purest form is a one-atom thick layer of carbon foil. In graphene, carbon atoms are densely packed in a regular (hexagonal) pattern. These hexagonal patterns are the basic structural element of related materials, including graphite, charcoal, carbon nanotubes and fullerenes. Graphene can also be thought of being an indefinitely large aromatic molecule. The hexagonal arrangement of carbon atoms in a single layer results in quantum mechanical effects that manifest at macro scale as extraordinary properties. Its interactions with other materials and with light in combination with its inherently two-dimensional nature produce unique properties, such as the bipolar transistor effect, ballistic transport of charges and large quantum oscillations. High-quality graphene is strong, light, almost transparent and an excellent conductor of heat and electricity.

applications. Given the state of graphene research at that time, the ‘Novoselov paper’ is to be considered a landmark event that transformed theoretical scientific knowledge into new technological applications. The ‘Novoselov paper’ linked theoretical considerations about graphene and its properties to experimental verification. The discovery presented in this paper proved that graphene could exist in ‘free form’⁵⁹ and provided a method to produce sheets of graphene, albeit of limited size. The resulting upswing in graphene R&D not only led to inventions and interesting applications, it also stimulated the application of modern theoretical physical concepts, much more advanced than those available to Wallace in 1947. The application of these new theories resulted in the discovery of previously unanticipated properties (Frenken, 2013). This renewed attention from theoretical physicists is reflected in the evolution of the cognitive structure as presented in graphene related publications. In view of the extended and expanding theoretical knowledge base, and the expected economic potential for graphene-based technologies we expect patent applications to appear soon, years rather than decades, after the breakthrough.

We searched for possible changes visible in bibliographic data related to the ‘Novoselov paper’ that can be used as indicators to help in the early stage identification of a publication as a potential breakthrough, and that can reveal the characteristics of this potential breakthrough. In this analysis we focus on the role the ‘Novoselov paper’ had on graphene-related developments, and the diffusion of the knowledge presented among researchers. We use bibliographic data to answer the question ‘Did this landmark paper by Novoselov et al. give rise to a new theoretical framework to explain the properties of graphene?’ We conclude that the ‘Novoselov paper’ describes a ‘charge’⁶⁰ breakthrough, and propose indicators that can be extracted from bibliographic data, and can be used to assist in identifying potential breakthroughs at early stage.

3.1.1 Justification for this study

Research on emerging scientific fields and technology forecasting is conducted from various perspectives, and with various goals in mind. In a preliminary study Adams (2005) concludes that for six research categories across the life and physical sciences there is a significant correlation between early (years 1–2) citation counts and later (years 3–10) citation counts. Andersen and Borup (2009) focus on improving the impact of foresight strategy processed in national research councils and research programs. A foresight model based on an alternative version of the Delphi-AHP approach to detect key areas in the Information Technology is proposed in (Bañuls and Salmeron, 2008). Bettencourt et al. (2008) shows that the evolution of emerging fields

⁵⁹‘Free form’ is to be understood as a 1-atom thick layer of graphene placed on a substrate. The only function of the substrate is to support and fix the graphene layer. The graphene in the layer behaves as if it were freestanding.

⁶⁰See the short description of Koshland’s Cha-Cha-Cha theory in footnote 61 on page 45.

in scientific disciplines is well described by population contagion models. In (Breiner et al., 1994) the outcomes of a joint Japanese German Delphi study are analysed to explain possible differences in the outcomes, and to understand the cultural influences on technology assessment. Chen et al. (2009) model transformative discoveries focussing on connections across structural holes in network representations of scientific knowledge in scientific discovery. Coates et al. (1994) reviewed and commented for 'Project 2025' all the science and technology forecasts for 1970 and any time forward they could find. Technology Forecasting (TF) and the changes TF underwent over the years are discussed in (Coates et al., 2001). Integrating multiples techniques to forecast emerging technologies by using bibliometric data and patent analysis as sources for historical data is discussed in (Daim et al., 2006). Hand (2009) addresses the challenges and difficulties involved in the use of the vast data sets used in forecasting. Julius et al. (1977) present an organisational model in which expert knowledge is used to judge at an early stage whether a discovery is expected to be of importance for new technology that will affect R&D in the pharmaceutical sector. Leydesdorff and Rafols (2011) use geographical and cognitive diffusion to analyse emerging topics. A model for a foresight process to identify scientific developments of strategic importance is presented in (Martin, 1995). Meade and Islam (1998) provides an overview of 29 different mathematical models that can be used in technology forecasting, and conclude "...The straightforward policy of identifying an appropriate model and then using it to generate forecasts was shown to be difficult, if not impossible, to put into practice ..." Mishra et al. (2002) propose the mapping on a common scale of characteristics of technology and techniques to match technique with the technology to be forecasted. Ponomarev et al. (2012, 2014b) use citation patterns in combination with statistical modelling to search and trace developments in potential breakthrough publications. A combination of a fuzzy Delphi method, analytic hierarchy process (AHP), and patent co-citation approach (ACP) to select emerging technologies is presented in (Shen et al., 2010). Small (1977) concludes that the citation picture of a specialty's development in science matches the perception of the specialists in the field. Tu and Seng (2012) propose two indexes, the novelty index (NI) and the published volume index (PVI) in order to detect emerging topics. Yoon and Park (2007) suggests combining morphology analysis, conjoint analyses and citation analysis of patent information to forecast new technologies. In (Young, 1993) nine different growth curve models are fitted onto various data sets in an attempt to determine the growth curve model that achieves the best forecasts.

We are looking for indicators in bibliographic data that signal, at early stage, the moment a publication of a potential breakthrough in science occurs. The precise point in time a breakthrough occurs often cannot, even in retrospect, be precisely isolated on a time line. We argue, however, that it is possible to gauge major breakthroughs retrospectively, within limits, by using bibliometric information to identify related landmark publications. Kosh-

land (2007) differentiates in his Cha-Cha-Cha theory⁶¹ scientific discoveries in three distinct classes on the basis of the nature of the discovery in relation to already existing scientific knowledge. Hollingsworth (2008, p.317) defines “A major breakthrough or discovery is a finding or process, often preceded by numerous small advances, which leads to a new way of thinking about a problem”. Our focus is on major role-changing discoveries, which we will call breakthroughs. Combining the ideas of Koshland and Hollingsworth enables us to differentiate between breakthrough discoveries and small advance discoveries, and opens up the possibility to classify breakthrough discoveries into the three categories defined by Koshland. In this study we interpret the aforementioned definition from Hollingsworth ‘...which opens up new ways for further physical research ...’ Kuhn (1962, Ch.VII,VIII) argues that crisis and theory change go hand in hand, and concludes that a crisis does not immediately lead to the replacement of the failing theory by a new one. ‘Challenge’ discoveries can be seen as revolutionary science in Kuhnian sense, and proliferate slow as result of reluctance to supersede an existing theoretical framework with a new one. ‘Charge’ discoveries on the other hand can almost instantaneously lead to new R&D activities as they solve scientific puzzles for which no immediate changes in the existing theoretical framework are needed. We therefore expect that a discovery leaves behind a pattern in bibliographic data that is specific for the type of this discovery.

The identification of breakthrough moments is relevant because they signal possible breaches (focus shifts) or even turning points in the R&D system. Martin (1995) points at the problems policy-makers and scientists have in selecting the most promising research areas and emerging technologies. Identifying at early stage publications of potential breakthrough discoveries could assist this selection process. Policy makers in industry or government could adopt their strategy based on the notion a possible breakthrough exists. We consider the moment patent applications that are closely connected to a research front start to appear signal that basic scientific knowledge from this research front becomes to be used for ‘real world’ applications. Geim and Novoselov wrote in the abstract of their publication “The rise of graphene” (Geim and Novoselov, 2007) “...the graphene ‘gold rush’ has begun.” The properties of graphene, forecasted by theoretical physical research, could only be exploited if freestanding graphene was isolated. The ‘Novoselov paper’ marks the start of the transformation of basic science into technological applications. This publication also leads to the application of more modern theoretical concepts than those available to Wallace and Boehm, which on their turn resulted in the prediction of other not yet foreseen properties.

We consider science-based innovations as innovations that are only made

⁶¹The three different types of scientific discovery according to Koshland (2007) are:

Charge discoveries, not necessarily simple, solve scientific problems that are quite obvious and fit within the at that moment existing theoretical framework.

Challenge discoveries are a response to an accumulation of facts or concepts that are unexplained by or incongruous with scientific theories of the time, ‘paradigm shifts’ (Kuhn, 1962).

Chance discoveries are those that are often called serendipitous. van Andel (1994) denotes discoveries of this type ‘unsought findings’.

possible by newly acquired scientific knowledge. For science-based innovation systems a long period between the development of the basic theoretical concepts and the first technological applications seems to be the rule rather than the exception. A number of researchers, including Jewkes et al. (1958), Isenson (1969), and Grupp and Schmoch (1992), conclude that the diffusion of basic scientific knowledge into technological applications often takes dozens of years. The accumulated theoretical knowledge is often released precipitately as a result of a single pivotal event (scholarly publication). Graphene is an example of such a science-based innovation. We see the ‘Novoselov paper’ as a pivotal publication and consider the graphene case to be particularly suited to our search for bibliographical indicators. The fact that there is one single event, the ‘Novoselov paper’, that describes the breakthrough discovery, results in a well-defined point in time that marks this event. According to the Nobel Prize committee (Nobel Prize Committee, 2010) this paper clearly marks the ‘graphene’ breakthrough. We discussed the ‘Novoselov paper’ and its effects on graphene research with professor Frenken, one of the researchers active in graphene research (see Table 3.5 for background information). We argue by analysing bibliographic data, applying Koshland’s criteria, and using Hollingsworth definition that the ‘Novoselov paper’ can be considered a typical example of a charge breakthrough. This conclusion is in line with information from the Nobel Prize committee, and the results of the discussion with Frenken (2013).

3.1.2 Hypothesis and research question

The accumulation of basic scientific knowledge of graphene took place over a period of almost 60 years. The ‘Novoselov paper’ showed a way to produce small sheets of freestanding graphene, in this way graphene became available for experimentation. We therefore expect the number of scholarly publications on graphene as well as the number of graphene related patent applications to rise significantly and simultaneously after a short ‘incubation’ period. A significant rise is anticipated because the breakthrough event removes the barrier for usage of the accumulated theoretical knowledge for applied scientific research. The expected upswing should, at first, coincide with the increase in the number of citations of the ‘Novoselov paper’ well before the Nobel Prize was awarded. We furthermore expect to find evidence in the bibliographic data authors and organizations not previously engaged enter the graphene arena.

Our hypothesis ‘Bibliographic data contains information that enables to identify and characterize at early stage publications of potential breakthroughs at the interface of science and technology’ is to be answered by the research question ‘Can we identify indicators in bibliographic information that can help in identifying and characterizing at early stage publications that describe a potential breakthrough at the interface of science and technology?’

To answer the research question the focus is on the following criteria to conclude if a typical publication contains a potential breakthrough that might

3.1 Introduction

result in a new technology. The criteria consist of two types, criteria 1 and 2 focus on very properties of a discovery, and the criteria 3, 4, and 5 on the impact of the discovery on the scientific research system:

1. the scientific discovery has the potency to evolve into a new technology
2. if the discovery introduces a new theoretical framework it is classified as a 'challenge' discovery, otherwise it is classified as a 'charge' discovery. Characterisation of a breakthrough is important as we think that breakthroughs leave behind specific patterns in bibliographic data that reflect the characteristics of the type of the breakthrough.
3. the publication is highly cited from the moment it is published. In this study we exclude publications with a delayed interest, the 'sleeping beauties' (van Raan, 2004)
4. the discovery leads to renewed interest for the theory underlying the discovery. This renewed interest is driven by the outcomes of experimental research related to the discovery
5. the discovery results in an above average influx of researchers and organisations new to the research field; the potential breakthrough gains much attention of scientist working in the same or adjacent areas, who want to participate in this new direction. The manner in which knowledge presented in a publication diffuses among scientist and organisations is an indication of the type of a discovery.

Hollingsworth (2008) provides a definition of a breakthrough, and in combination with Koshland's Cha-Cha-Cha theory this should make it possible to distinguish between charge discoveries and challenge discoveries. Consulting an expert on the subject helps in validating the findings, and especially to conclude if a paradigm shift emerged from the discovery.

We analyse bibliographic information related to the 'Novoselov paper' to learn if these criteria can be used to base indicators on that can be of help in early stage identification and characterization of potential breakthroughs at the interface of science and technology.

3.1.3 Data and information sources

Scholarly publications

We used Thomson-Reuters Web of Science database (WoS) as source for bibliographic information on scholarly publications. The WoS contains, according to Thomson-Reuters website, over 55 million records with bibliographic data for scholarly publications from journals (> 18,000), conference proceedings, and books in the sciences, social sciences, and arts and humanities. For a detailed analysis of the evolution of graphene research we collected, from the in-house version of the WoS (TR/CWTS WoS), for the relevant publications terms occurring in titles and abstracts to create co-occurrence maps. We used the date of publication to put scholarly publications on the time scale. The publication date used in the WoS is the 'date' on the front cover of the

journal.⁶²

As long as graphene was not available as freestanding substance research was closely linked to research on graphite; graphene was hardly seen as a separate field of research. To select relevant publication from the WoS we used the topic⁶³ 'graphene'. This approach resulted in a data set containing information for 18,499 scholarly works⁶⁴ for the period 1990–2012.

Patent publications

We used the April 2012 version of PATSTAT⁶⁵ to collect the relevant patent data. To select patent publications we relied on graphene-specific patent classification codes assigned to the patent publications. The use of patent classification codes is to be preferred over keyword searches; as skilled patent examiners assign these classification codes to patent documents. Widely in use is the International Patent Classification system (IPC). Appropriate classification codes from the IPC are assigned to all patent applications. The IPC however does not contain graphene specific codes to date. A more fine-grained classification system compatible with the IPC was in use at the European Patent Office (EPO) until 2013⁶⁶. This system called the European Classification system (ECLA) was accompanied by an even more specific classification system called *In Computer Only* (ICO). Both ECLA and ICO contain several graphene-specific classification codes⁶⁷. Patent examiners at the EPO assigned the moment a patent application was processed appropriate ECLA and ICO codes.

Graphene is considered a material with a bright technological future, and is expected to generate large revenues. Therefore we expect that the vast majority of graphene related inventions would end up as patent application at the EPO as one of the major patent offices in the world. The drawback in using ECLA and ICO codes is that no such code is assigned to a patent application until the moment it is processed by an examiner at the EPO. We compensate for this systemic error by additionally including in our data set bibliographic information for patent documents of which the abstract or the

⁶²A more accurate approximation for the date of a discovery would be the date a manuscript was submitted for the first time to a publisher. This submission date is not available in the databases and could therefore not be used. The publication itself is not necessarily the first attempt to publish the discovery resulting in an even more imprecise time stamp for the discovery.

⁶³ According to Thomson-Reuters information terms for a topic are extracted from the title, the abstract, the author keywords of a publication, and from KeyWords Plus[®]. 'KeyWords Plus are words and phrases harvested from the titles of the cited articles.'

⁶⁴ In this study we focus on publications containing original scientific work, and therefore only publications of the WoS types 'article' and 'letter' are used in the analysis

⁶⁵EPO Worldwide Patent Statistical Database (PATSTAT)

⁶⁶The Cooperative Patent Classification (CPC), which is the result of cooperation between the EPO and US Patent and Trademark Office (USPTO) replaced in 2013 the ECLA, ICO, and the classification of the USPTO (US Patent Classification).

⁶⁷We used ECLA and ICO codes C01B31/04H, H01F41/42, H01F42/44, H01F10/00C, H01L29/16G, M01B204/00, M01B204/02, M01B204/06, M01B204/20, T01F10/00C, T01L29/16G

3.2 The graphene breakthrough

title contains the phrase ‘graphen*’. Other search terms were tested but inspection of the results learned that these did not reveal additional relevant publications without introducing non-relevant documents. We are confident that the dataset covers the patent publications in an adequate way.

An intrinsic effect of the patent system is that inventions often lead to multiple equivalent patent publications in several jurisdictions⁶⁸. To count ‘inventions’ we grouped equivalent patent documents together in DOCDB-patent families; such a patent family represents one invention. The earliest filing date of all patent applications within a patent family is used as approximation for the date of the invention. The resulting data set contained 2,083 patent families applied for in the period 1990–2011. The data set does not fully cover patent filings from Autumn 2010 onwards due to the secrecy period of 18 months, that is part of the patenting procedures, and the use of the April 2012 release of PATSTAT.

Citations in patents reference other patents as well as other, for instance scholarly, publications. Especially patent publications published by the United States Patent and Trademark Office (USPTO) contain a substantial list of these front-page⁶⁹ citations. Other patent offices usually publish only citations that occur in ‘search reports’; the citations mentioned in these reports constitute only a small portion of the citations in the patent documents. Especially in the case of technologies considered very promising, like graphene, there is often at least one publication from the USPTO present in the patent family. Of all DOCDB families in the PATSTAT database approximately 20% contain at least one USPTO publication; for the graphene patent publications in our dataset this is the case in 41% of the patent families. In the analysis all relevant bibliographic data were taken into account. The use of patent families solves, at least in part, the problems with references in patent publications to scholarly literature.

3.2 The graphene breakthrough

Section 3.2.1 focuses on the evolution of the graphene field, and uses absolute numbers of scholarly publications and patent applications to visualize it. Breakthrough publications are expected to stimulate research, and therefore result in a steep increasing number of related publications. To analyse the impact of the ‘Novoselov paper’ over time the share of scholarly publications and patent publications citing this publication are calculated for each year. It is expected that breakthrough publications be among the publications with a high citation count. In subsection 3.2.2 we compare the number of citations the ‘Novoselov paper’ received during the first 12, 24 and 36 months with the numbers received by graphene publications published in 2004 and 2005, and by the publications from Volume 306⁷⁰ of Science. According to

⁶⁸Patent rights are nowadays national rights granted by a sovereign state

⁶⁹In the US the list of citations appear on the front-page of the patent publication

⁷⁰The volume in which the ‘Novoselov paper’ was published

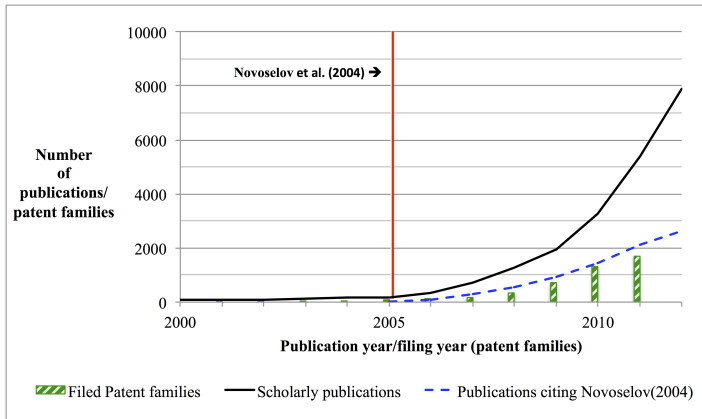


Figure 3.1: Graphene related scholarly publications and DOCDB patent families (2000–2012)

Koshland ‘Challenge’ discoveries lead to changes in the theoretical framework. The publications that cite de ‘Novoselov paper’ were classified into six categories; ranging from discovery science to applied science. We calculated (Section 3.2.3) for each year in the period 2005–2012 for every category the absolute numbers of publications, and the share of total publications in that year. A drop in the number of publications in the category ‘discovery science’ while at the same time the numbers for the other categories increase is an indication that the scientific knowledge from the publication becomes used in applied areas. The balance between basic research and applied research is discussed in Section 3.2.4. The publication of a breakthrough discovery stimulates researchers to become active in the new area, and start citing in their publications the breakthrough publication. In Section 3.2.5 the first occurrence of first authors, and their affiliations, that cite the ‘Novoselov paper’ in their publications is used to approximate this influx. The possible effect of the ambiguity of names is discussed in Section 3.2.6.

3.2.1 The evolution of the graphene field

Figure 3.1 presents trends in absolute numbers for graphene-related scholarly publications, publications citing the ‘Novoselov paper’, and for patent families. After the publication of the ‘Novoselov paper’, the ‘graphene field’ shows an upswing in R&D visible as a sharp increase in the number of publications later accompanied by a remarkable rise in patent applications. The figure shows furthermore the number of scholarly publications citing the ‘Novoselov paper’. The share of these publications topped in 2009 at 49%, and since declines as is shown in Table 3.1.

It is known that certain events, such as the award of a Nobel Prize, can have a significant influence on the number of citations that publications of laur-

3.2 The graphene breakthrough

eat authors receive. In this study this effect plays no role as the focus is on citations made well before the Nobel Prize Physics was awarded to Geim and Novoselov in 2010. The last two columns in Table 3.1 show that citations in patent publications to the ‘Novoselov paper’ are present since 2006⁷¹. The patent data for 2010 and 2011 is incomplete as result of the secrecy period, which is part of the patenting process. Citations made by patent applicants are comparable in nature with citations in scholarly publications. In most cases the applicant of a patent, as is shown in this table, added the citations to the ‘Novoselov paper’. These citations indicate the existence of a link between science (scholarly publications) and technology (patent publications). The number of patent families citing the ‘Novoselov paper’ is low in comparison with the number of ‘graphene’ patent families that have been filed. The discovery in the ‘Novoselov paper’ showing the method to isolate graphene as a substance so it could be used in experimental research influenced, directly or indirectly, much graphene related research.

3.2.2 Has the ‘Novoselov paper’ been a highly cited paper from the moment it was published?

The low number of citations in patent applications suggests that to the majority of graphene related inventions the discovery is not seen to be of particular significance, and as patent applicants are not obliged to cite previous literature in their patents it is therefor not cited. Further evidence for the limited importance of this paper for patenting, is given by the fact that the patent families in the dataset contain in total 2,462 references to scholarly publications; only a small share (3%) are citations to the ‘Novoselov paper’; these citations originate predominantly from patent applicants. Patent examiners usually only cite publications that are relevant to the granting decision they have to take, and document them in search reports. By using patent families of which, at least in this case, a large number (41%) contain a USPTO patent publication a bias towards citations originating from the patent examiners is largely prevented. The reason for this is that in the US patent applicants are obliged to cite all relevant patent and non-patent literature on which their invention is based. The large influence of the ‘Novoselov paper’ on the development of graphene related R&D is shown by the share of scholarly publications citing it. The share of graphene publications citing that paper topped in 2009 with 46%, and decreased since; the absolute number of citations is still rising. As the graphene research field matures research on graphene digresses from the landmark publication. We conclude that the ‘Novoselov paper’ is at the interface of science and technology as it is referenced in patent publications. The direct influence on the development of graphene technology seems to be minor as the paper is cited in only a small number of graphene related patents. This should not come as a surprise as the discovery in the ‘Novoselov

⁷¹Inspection of the patent file for the patent application from 2004 citing the ‘Novoselov paper’ leads to the conclusion that at a later stage in the patenting process this citation was added; most probably in 2006.

Table 3.1: Scholarly publications (articles, letters), patent filings and citation to the ‘Novoselov paper’

Year	Scholarly publications on the topic graphene			Graphene DOCDB patent families	
	Number of publications	Publications citing the ‘Novoselov paper’	share of scholarly publications citing the ‘Novoselov paper’	Citations to the ‘Novoselov paper’	(made by the patent applicants)
2000	81	.	.	29	.
2001	90	.	.	39	.
2002	104	.	.	38	.
2003	130	.	.	52	.
2004	151	.	.	49	1 ^b
2005	182	10	5%	46	0
2006	323	81	25%	88	2
2007	703	272	39%	142	1
2008	1,270	553	44%	318	9
2009	1,933	880	46%	650	8
2010	3,246	1,402	43%	614 ^c	13
2011	5,363	2,024	38%	226 ^c	2
2012	7,885	2,543	32%	.	.

^a ‘.’ indicates that data is unavailable.

^b this citation was most probably added in 2006 (see also footnote 71 on page 51)

^c the data is incomplete as result of the secrecy period that is part of the patenting. procedures. Patent families filed in 2010 are for a greater part covered by the PATSTAT release that was used; the patent families filed in 2011 are covered for only a small part.

paper’ is in producing, isolating, identifying and characterizing graphene. Inspection of the patent publications shows that most inventions focus on the application of graphene and not on producing graphene.

In order to assess whether the ‘Novoselov paper’ is a highly cited paper we compared the number of citations it received with the numbers for all other graphene publications⁷² from 2004 and 2005. We computed the share of the publications that received a certain number of citations within 12, 24, and 36 months after publication. The results are shown in Figure 3.2. The Y-axis in this figure represents the share of the publications that received at most the number of citations of the corresponding point on the X-axis. The position of the ‘Novoselov paper’ is indicated on each of the curves by opaque (red) disks. This figure and also Table 3.2 illustrate that among the graphene publications from 2004 and 2005 the ‘Novoselov paper’, from the moment it

⁷²We included only publications with document types research article and letter

3.2 The graphene breakthrough

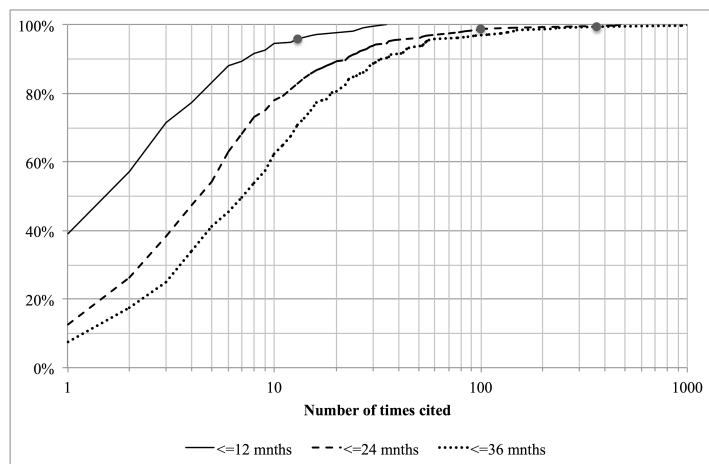


Figure 3.2: Cumulative distribution of graphene related publications from 2004 and 2005 based on the number of citations received during 12, 24, and 36 months after publication. The opaque (red) disks indicate the ‘Novoselov paper’

was published, is one of the most cited. Based on the number of citations received in 36 months (Table 3.2) the ‘Novoselov paper’ belongs to the top 1% of publications from Volume 306 of Science, the volume it appeared in. The publication in Science, an esteemed journal, with its wide audience might have contributed to the fact that it received from the moment it was published a considerable number of citations. By the end of the year 2013 the ‘Novoselov paper’ was cited 11,623 times in research articles, and 575 times in review⁷³ publications. We furthermore compared the number of citations this landmark paper received with the number of citations received by the 318 publications published in the same volume (306) of Science. The last column of Table 3.2 shows the results. Based on citation counts for the first 24 months the Novoselov paper was in all cases among the top 5% of publications.

3.2.3 Did the ‘Novoselov paper’ introduce a new paradigm?

The ‘Novoselov paper’ did not introduce a new paradigm as was implicitly mentioned in (Nobel Prize Committee, 2010), and explicitly in (Frenken, 2013). To check if bibliographic data contains information to answer this question we used the classification scheme presented Table 3.3 to assign one of six science levels to publications⁷⁴. The methodology and background of this

⁷³In this paper the classification of a publication is the classification as it appears in the Web of Science database.

⁷⁴The science levels are assigned to journals on the basis of the addresses of the authors of the publications in a journal. Based on the addresses the affiliation of an author is classified as

Table 3.2: Citations to the ‘Novoselov paper’ in the first 12, 24, and 36 months after publication, the ranking among all graphene publications from 2004 and 2005, and the ranking among all publications from the same Volume (306) of Science on the basis of the number of citations.

Citation window (time lag between publication and citation)	Number of citations to the ‘Novoselov paper’	Ranking among all graphene publications from 2004 and 2005	Ranking among all publications of Science Volume 306
≤ 12 months	18	top 4%	top 30%
≤ 24 months	105	top 3%	top 5%
≤ 36 months	373	top 1%	top 1%

Table 3.3: Journal classification science levels

Category number	Science level
1	Discovery Science
2	Industrially Relevant Science
3	Science-Based Technological Development
4	Clinically Relevant Science
5	Science-Based Clinical Practice
6	Industrial/Medical Development

classification are presented in (Tijssen, 2010). The classification ranges from ‘Discovery Science’ to ‘Applied science’. Individual publications inherit the science level assigned to the journal in which they are published. Only publications with science levels ‘Discovery science’, ‘Industrially relevant science’, ‘Science-based technological development’, and ‘Industrial/medical development’ appear in the document set. The class ‘Discovery science’ is assigned to the ‘Novoselov paper’ as this is the science level assigned to the Science journal.

To answer the question “Did the ‘Novoselov paper’ introduce a new paradigm?” we analysed the evolution of the number of publications citing the ‘Novoselov paper’ that are assigned to the different science levels. ‘Discovery Science’ is the category that encompasses theoretical research. Figure 3.3 shows the absolute numbers of publications citing⁷⁵ the ‘Novoselov paper’ disaggregated into science levels. Publications in the category ‘Industrially Relevant Science’ dominate the picture from the beginning.

In Figure 3.4 the shares of the different science levels of the citing publications per year are shown. The share of publications citing the ‘Novoselov paper’ from the category ‘Science-based technological development’ increased

university, hospital, or company. The distribution of the three classes determines the science level assigned to a journal

⁷⁵In order to get reliable results (Tijssen, 2010) not all journals, and hence not all publications, have a science level assigned. This results in numbers of publications that are somewhat lower than those presented in Table 1.

3.2 The graphene breakthrough

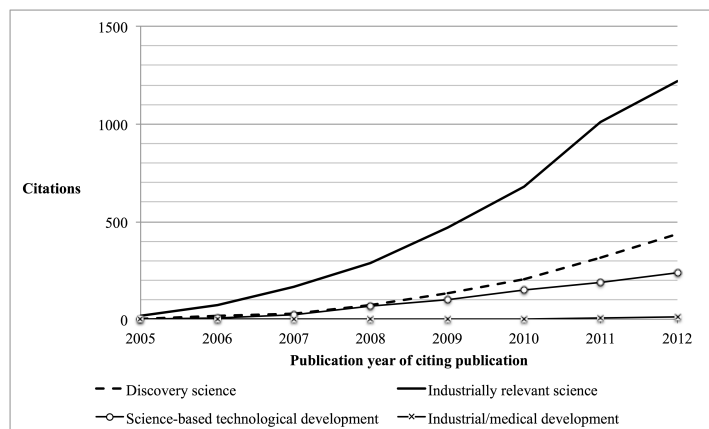


Figure 3.3: Scholarly publications citing the ‘Novoselov paper’ disaggregated into science levels

from 2005 until reaching a maximum in 2008; from then on it gradually declines. Since 2005 the share of publications in the ‘Discovery science’ category citing the ‘Novoselov paper’ increases at a slow pace; in 2007 it dropped from 20 to 17% and is increasing since. This is in accordance with the notion that theoretical physicists became again interested in graphene and started to apply more recent theoretical approaches. It is also in line with the discussion in the next Section 3.2.4. From the beginning the majority of the publications belong to the category applied science, especially ‘Industrial Relevant Science’. In case the ‘Novoselov paper’ would have introduced a new theoretical framework we expect, at least during period immediate following the publication, dominance in absolute and relative numbers of publications from the ‘Discovery Science’ category citing it; this is not the case. We therefore conclude that the bibliographic data shows evidence that the ‘Novoselov paper’ did not introduce a new theoretical framework.

3.2.4 Balance between basic research and applied research

To analyse the evolution of graphene related research we extracted terms from the abstracts and titles of all graphene related papers in the document set. (Winnink and Tijssen, 2015, supplementary material) visualizes for the period 2004–2012 the cognitive structure of graphene research in every year by creating co-occurrence map⁷⁶. For each year the co-occurrence maps are the result of combining terms from the abstracts and titles of all graphene publications up-to that year. Development and shift of focus in graphene research should be visible as changes in these co-occurrence maps⁷⁷ over time.

⁷⁶This information is integrated in this thesis as Section 3.A on page 64.

⁷⁷These pictures were presented before at the STI 2012 conference (Winnink, 2012)

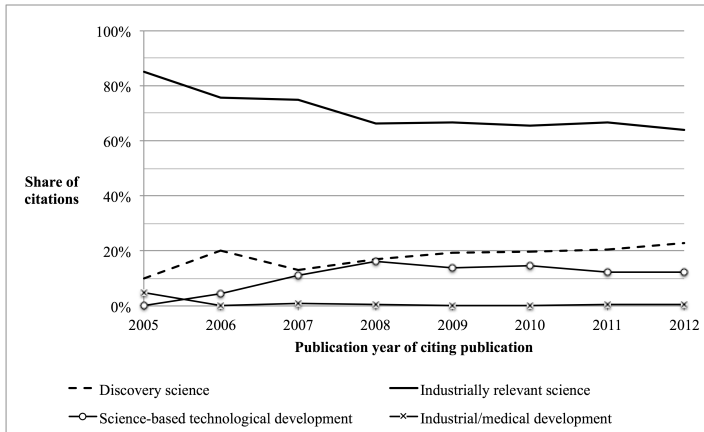


Figure 3.4: Publications per science category as share of the total number of scholarly publications citing the ‘Novoselov paper’

We used the VOSviewer⁷⁸ for the visualisation. For all pictures the same settings for the visualisation parameters were used to assure that if changes could be observed these changes are caused by alterations in the cognitive structure, and are not the result of different parameter settings. A dichotomy in two areas of more or less equal intensity becomes manifest in 2007. From 2007 onwards the pictures show a more intense area on the right hand side that is related to basic theoretical research, and a second area composed of terms related to applied scientific research on the left. We expect this constellation in which the two areas are in ‘balance’ to be meta-stable; as time proceeds and graphene becomes more mature we expect this balance to change in the direction of applied scientific research. The pictures shows that applied scientific research became more important but at the same time basic (theoretical) research did not diminish. The data suggests that discovery science supports the further development of graphene. This observation is in line with the information presented in Figure 3.4, and the outcomes of the discussion with Frenken (Section 3.5) on the question “What causes the dichotomy between applied science and theoretical science that becomes apparent in 2007–2008 in the co-occurrence maps of terms from the abstracts and the titles of graphene publications, and that evolves into a (meta) stable configuration?”

The evolution shown in the pictures in Section 3.A on page 64 is also in line with Frenken (2013) stating that the ‘Novoselov paper’ was at the basis of research that proved existing predictions of the properties of graphene to be true, and stimulated the application of more recent theoretical insights that led to the prediction of new properties.

⁷⁸<http://www.vosviewer.com>

3.2.5 The inflow of researchers and organisations in graphene research as result of the ‘Novoselov paper’

In some cases discoveries stimulate other researchers to do research related to the discovery, and to publish the outcomes of their research. Our focus is on the diffusion among scientist and research organisations of the knowledge presented in a publication. We question if characteristics of this diffusion process can be used to identify breakthroughs at early stage. To do this we track scientists that become active in graphene research, and that cite the ‘Novoselov paper’ in their publications. To measure this influx of researchers we focus on first authors of publications and their affiliations⁷⁹ and define two quantities *New First Authors (NFA)* and *New First main Organisations (NFO)*. These quantities measure the number of first authors and their affiliations that become active within a time interval, and who have not been mentioned before as first author or first organisation on a publication citing the ‘Novoselov paper’. The set of distinct authors at time T is presented as $N_{A,T}$, and the set of distinct main organisations at time T as $N_{O,T}$. For a period ranging from t_1 until t_2 *NFA* is defined as $N_{A,t_2} - N_{A,t_1}$ and *NFO* as. Based on these two quantities we define the relative mutation of *NFA* (*RelMutNFA*) as $\frac{N_{A,t_2} - N_{A,t_1}}{N_{A,t_1}}$, and *NFO* (*RelMutNFO*) as $\frac{N_{O,t_2} - N_{O,t_1}}{N_{O,t_1}}$. These two quantities measure *NFA* and *NFO* in relation to the number of distinct authors and distinct organisations that are already active at the beginning of the interval (t_1). The ratio of *NFA* and *NFO* (*NFANFOratio*) indicates if authors and organisations become active at the same pace, and is defined as $\frac{N_{A,t_2} - N_{A,t_1}}{N_{O,t_2} - N_{O,t_1}}$. We define two more quantities *NFAcitsratio* as $\frac{Ncits_{t_2} - Ncits_{t_1}}{N_{A,t_2} - N_{A,t_1}}$, and *NFOcitsratio* as $\frac{Ncits_{t_2} - Ncits_{t_1}}{N_{O,t_2} - N_{O,t_1}}$. These quantities compare the number of authors, and organisations that become active with the increase of the citation count in the same period. If the increase of the citation count is the result solely of authors or organisations entering the field the ratio is close to 1.0. When already active authors and organisations continue producing citations the ratio is above 1.0.

We compared the values for publications⁸⁰ citing the ‘Novoselov paper’ with the values for 318 publications from Volume 306 of Science, and the 66,073 publications from October 2004 — the month of the ‘Novoselov paper’ was published — to find out if the proposed quantities have a discriminating effect on the publications. In all cases citations within 24 months after the publication of a cited document are taken into account. First for each cited publication the values were calculated separately. The distributions of the

⁷⁹First author = the author first mentioned on a publication, first organisation = the affiliation of the first author of a publication. Ambiguity in names and the effects for the analysis are discussed in subsection 3.2.6. Authors and affiliations are counted only the first time they publish a publication citing the ‘Novoselov paper’; we consider this moment in time the moment the author and his or her organisation enters graphene research stimulated by the ‘Novoselov paper’

⁸⁰Only publications covered in TR/CWTS WoS of types article and letter are taken into account to focus on publications containing original research

Table 3.4: For articles and letters citing the ‘Novoselov paper’ the quantities NFA , NFO , $RelMutNFA$, $RelMutNFO$, $NFANFORatio$, $NFAcitsratio$, and $NFOcitsratio$ are compared with the values obtained for the 318 publications from Science Volume 306, and the 66,073 publications published in October 2004. Citations within 24 months after the publication of a cited document are considered.

Quantity	Publication(s)		Percentile borders						
			0%	25%	50%	75%	90%	95%	99%
NFA	‘Novoselov paper’	52.0							
	Science Volume (306)		0.0	3.0	0.5	19.0	30.0	41.7	69.0
	October 2004		0.0	0.0	1.0	3.0	5.0	7.0	15.0
NFO	‘Novoselov paper’	41.0							
	Science Volume (306)		0.0	3.0	8.0	14.0	21.3	29.7	48.0
	October 2004		0.0	0.0	1.0	2.0	5.0	6.0	13.0
$RelMutNFA$	‘Novoselov paper’	5.8							
	Science Volume (306)		0.0	1.0	1.7	2.4	4.0	5.1	8.1
	October 2004		0.0	0.0	0.0	1.0	2.5	4.0	7.0
$RelMutNFO$	‘Novoselov paper’	4.6							
	Science Volume (306)		0.0	0.8	1.3	2.2	3.4	4.3	9.3
	October 2004		0.0	0.0	0.0	1.0	2.0	3.0	6.0
$NFANFORatio$	‘Novoselov paper’	1.3							
	Science Volume (306)		0.0	1.0	1.2	1.4	1.7	1.8	2.2
	October 2004		0.0	0.0	1.0	1.0	1.3	1.6	2.3
$NFAcitsratio$	‘Novoselov paper’	1.4							
	Science Volume (306)		1.0	1.0	1.0	1.2	1.3	1.4	1.7
	October 2004		1.0	1.0	1.0	1.0	1.5	2.0	3.0
$NFOcitsratio$	‘Novoselov paper’	1.7							
	Science Volume (306)		1.0	1.1	1.3	1.7	2.0	2.3	3.0
	October 2004		1.0	1.0	1.0	1.3	2.0	2.0	4.0

3.3 Expert opinion

values for the various quantities were computed from the values obtained for the individual publications. Table 3.4 shows that for the ‘Novoselov paper’ *NFA*, *NFO*, *RelMutNFA*, and *RelMutNFO* are in the top 5% and in a number of cases in the top 1%. The *NFANFOratio* for the ‘Novoselov paper’ is in the top 50% for the publications in Science, and in the top 10% of the publications from October 2004. *NFAcitsratio* for the ‘Novoselov paper’ is in the top 5% of the publications in Science Volume 306, and in the top 11% for the publications from October 2004. *NFOcitsratio* is in the top 25% the publications in Science Volume 306, and in the top 13% for the publications of October 2004.

3.2.6 *Influence of ambiguity of author and organisation names*

The names of authors and organisation in the Web of Science database are known to be ambiguous in several cases. This ambiguity complicates the identification of unique authors and unique main organisations. In the computations we used the names as they appear in the database (TR/CWTS WOS) without any further disambiguation. To check for the effects of ambiguous names we took a random sample of 1,000 publications out of the 66,073 publications from October 2004. For each publication in this sample only the publications citing the ‘Novoselov paper’ within 36 months were considered; the first authors names and their affiliations were collected, and checked for obvious data errors. The author names did not reveal any obvious errors. In 18 cases the main organisation name was not available in the database, and in 10 cases the main organisation names contained obvious errors such as misspellings or inconsistent use of abbreviations. These latter errors could be corrected easily. A further check was done on the name of the first author in combination with the main organisation name. If a combination was identical for publications citing the same document the names were considered to be unambiguous for the cited publication. The occurrence of multiple organisations for the same first author citing the same publication was found in 341 cases. In these cases more than one author might be involved; other options such as authors with multiple affiliations, and authors switching from one organisation to another during the period are possible. We conclude that in 341 (9%) cases the author names might be ambiguous. No cases were found for which the group of articles and letters citing the same publication contained an ambiguous main organisation name.

3.3 Expert opinion

We discussed preliminary findings of this study with Frenken (2013), a senior scientist active in graphene research. Table 3.5 briefly presents information on Frenken’s professional background. The discussion focused on three questions.

1. *Did the 'Novoselov paper' introduce new theoretical concepts, a paradigm shift?*

As motioned in (Boehm et al., 1962; Nobel Prize Committee, 2010; Wallace, 1947) presented the theoretical background for the graphene breakthrough published in 2004. The discovery in the 'Novoselov paper' is about producing, isolating, identifying and characterizing graphene and as such does not introduce a paradigm shift.

2. *What causes the dichotomy between applied science and theoretical science that becomes apparent in 2007–2008 in the co-occurrence maps of terms from the abstracts and the titles of graphene publications, and that evolves into a (meta) stable configuration?*

The observed dichotomy signals the renewed interest from theoretical physicists for graphene leading to the application of much more modern theories than those available to Wallace in 1947 or Boehm in 1962. The moment experiments showed that the predicted properties for graphene existed in reality, theoretical physicist became again interested and started to apply more recent theoretical approaches. These new approaches led to the prediction of new properties. The consequence is that publications focused on theoretical aspects of graphene kept appearing and together with the up rise of publications on applied graphene research led to the division of graphene publications in two main areas.

3. *What is the challenge in contemporary graphene research?*

The current challenge is to produce sheets of graphene that are substantially bigger than the small pieces, roughly 1 cm^2 , that can be produced at this moment.

3.4 Discussion and concluding remarks

The goal of this study to prove our hypothesis 'Bibliographic data contains information that enables to identify and characterize at early stage publications of potential breakthroughs at the interface of science and technology' to be true. We analysed bibliographic data for the publication (Novoselov et al., 2004) as a well-known example of a discovery that is generally considered a breakthrough. In our research we differentiate breakthroughs using Koshland's Cha-Cha-Cha theory. We do this as we think that breakthroughs leave behind patterns in bibliographic data that reflect the characteristics that are typical for the type of the breakthrough. This study concentrates on five statements to uncover bibliographic information that could help in identifying at early stage publications containing a potential breakthrough discovery.

1. The scientific discovery has the potency to evolve into a new technology discovery. We view patent publications as representatives of technological developments. Scholarly publications cited in patents, especially those cited by patent applicants, show direct links between science and technology. The fact that the 'Novoselov paper' is cited in patents shows that this publication links science and technology. Figure 3.1 shows fur-

thermore that from 2006 onwards the number of ‘graphene’ related patent filings increased following with a time lag the up rise in the number of graphene related scholarly publications, and especially the publications citing the ‘Novoselov paper’.

2. If the discovery does introduce a new theoretical framework it is classified as a ‘challenge’ discovery, otherwise it is classified as a ‘charge’ discovery. The subject of the ‘Novoselov paper’ is obtaining and identifying graphene, and it is therefore a ‘technical’ publication important to obtain this almost mythical material to do experiments on. Our discussion with professor Frenken confirmed that the ‘Novoselov paper’ did not introduce a new theoretical framework. By assigning science levels to the publications citing the ‘Novoselov paper’ and analysing the evolution of the number of publications assigned to these science levels supports this view. This characterizes the discovery in the ‘Novoselov paper’ as a charge breakthrough as breakthroughs of this type, by definition, do not introduce paradigm shifts.
3. The publication is highly cited from the moment it is published. As shown in Figure 3.2 and Table 3.2 the ‘Novoselov paper’ from its publication received many citations. It received these citations well before the Nobel Prize Physics 2010 was awarded.
4. The discovery leads to renewed interest from theoretical physicists. The evolution of co-occurrence maps (Winnink and Tijssen, 2015, supplementary material) based on the terms used in the publications citing the ‘Novoselov paper’ revealed that from 2007 onwards the terms clearly cluster in two areas. Both areas are in ‘balance’; one cluster denotes basic research and the second cluster focuses on applied research. The results presented in Figure 3.4 support this conclusion that is also in line with information from Frenken. The conclusion is that theoretical physicists regained interest in graphene around three years after the publication of the ‘Novoselov paper’. This renewed interest was not the result of a paradigm shift, but was driven by the outcomes of experimental research that confirmed predictions already made by Wallace (1947) and (Boehm et al., 1962). Theoretical physicists became interested and applied more recent theoretical insights to the ‘graphene issue’ and came up with new predictions (Frenken, 2013).
5. The discovery results in an above average influx of researchers and organisations new to the research field. Using the quantities NFA , NFO , $RelMutNFA$, $RelMutNFO$, $NFANFOratio$, $NFACitsratio$, and $NFOcitsratio$ we try to measure the diffusion of the knowledge presented in a particular publication among researchers and research institutes. This study indicates that these quantities have discriminating power, and might be usefull to classify publications between those containing a potential breakthrough discovery and those that do not. For most quantities the ‘Novoselov paper’ gets high values. The use of these measures presupposes the unambiguity of names of authors and research organisations; in general this is not the case. We checked the

ambiguity of names and found that it plays a minor role. The reason for this is that the probability for a name occurring more than once on publications that cite the same document in a short period, 36 months or less, is limited. During testing we found that 9% of the names might be ambiguous.

In this study we used (Novoselov et al., 2004) as typical example of a breakthrough at the interface of science and technology. Citation relations form the basis in our analysis. The analyses of the effects of 'Novoselov paper' on graphene R&D shows that it is a typical example of a 'charge' breakthrough, and that no paradigm shift is involved. We propose a set of measures to be derived from bibliographic data that could help in identifying at early stage publications that contain a potential scientific breakthrough. The proposed measures have discriminating power. Ambiguity of names showed not to be an issue for these measures due to the short time period used in the analyses in combination with the focus on citations to one particular publication -in this case the 'Novoselov paper'. The importance of identification of breakthroughs in science at early stage is that it could aid policy makers, funding organisations, and companies to timely prioritize resources for R&D.

The scientific community is, in this study, seen as an analytical instrument that evaluates scientific discoveries, and as result generates 0 or more citing publications. The bibliographic information for these citing publications reflects the opinion of the scientific community on the cited publication. The measures we derived in this study suggest that they can be used to differentiate publications into those that probably contain a potential breakthrough and those not. We do realize that this case study is an analysis of only one well-known breakthrough; the measures we propose need to be validated for general applicability; such a study is foreseen as a follow up. In this follow up study we will test the outcomes of this and other studies on random sets of publications.

Our intention with this study was to find information in bibliographic data that might be the basis for indicators that could help in identifying and characterizing at early stage potential breakthrough discoveries at the interface of science and technology. Not addressed in this study is the fact that discoveries considered at first a breakthrough might at a later stage prove not to be. The importance of identification of breakthroughs in science at early stage is that it could aid policy makers, funding organisations, and companies to timely prioritize resources for R&D.

3.5 Acknowledgements

The authors kindly thank Prof. J.W.M. Frenken, Leiden Institute of Physics - Leiden University, and Prof. A.F.J. van Raan, Centre for Science and Technology Studies - Leiden University, for their useful comments on an earlier version of this manuscript. We also like to thank the anonymous reviewers for their valuable comments during peer review of this manuscript.

Table 3.5: Information on the professional background of professor Frenken

Joost Frenken is a Full Professor at the Physics Department (LION) of Leiden University, The Netherlands, where he is head of the Interface Physics Group.

... Central to the research of Joost Frenken is the fundamental understanding of the dynamic aspects of surfaces and interfaces and their role in relevant, natural or industrial processes under practical conditions. Topics of interest include surface diffusion, crystal growth, surface phase transitions, catalysis, and friction. For tailor-made measurements in each of these areas, Frenken's research group has developed a variety of special-purpose scanning probe microscopes as well as dedicated instrumentation for surface x-ray diffraction.

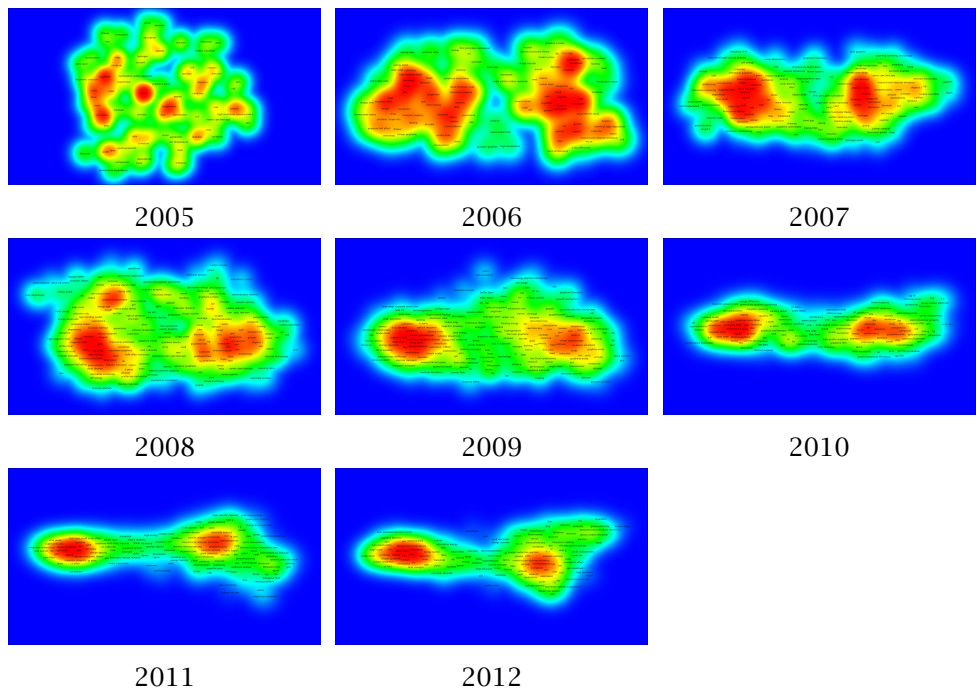
... Frenken has been (co)-initiator of two spin-off companies. Leiden Probe Microscopy BV markets commercialized versions of the research equipment developed in Frenken's research group, with an emphasis on advanced scanning probe and x-ray diffraction instrumentation. Applied NanoLayers BV is a recent spin-off, specialized on the production of high-quality graphene and boron nitride monolayers.

(<http://ip.physics.leidenuniv.nl/index.php/people/frenken/9-cat-members/51-frenken-bio>) (2013-12-23)

As of January 1st 2014, Joost Frenken will be the director of a new scientific research center. This research facility focuses on physical and chemical processes that are crucial to extreme UV lithography. This new research center, ARCNL, is a collaboration between ASML, FOM, NWO and the two universities in Amsterdam (University of Amsterdam and the VU). (<http://ip.physics.leidenuniv.nl>) (2013-12-23)

Appendix 3.A **Supplementary information — co-occurrence term maps**

Table 3.6: Evolution of the knowledge base about Graphene using co-occurrence term maps for the period 2005–2012 (Winnink and Tijssen, 2015). The graphs were also presented at the STI conference in 2012 (Winnink, 2012). The maps are based on the abstracts of Graphene related scholarly publications up to the end of the indicated year.



Chapter 4

Case study #3: Ubiquitin

The content of this chapter is equivalent to the manuscript published as:

Winnink, J.J., Tijssen, Robert J.W., and van Raan, A.F.J. (2015). Theory-changing breakthroughs in science: the impact of research teamwork on scientific discoveries. *Journal of the Association for Information Science and Technology*, (DOI:10.1002/asi.23505).

The only differences between the published article and the text in this chapter are of typographic nature to assure that the layout is in line with the typographic design of this PhD thesis. The information presented in section 'Appendix 4.A Supplementary material' on page 84 is also available on the website of the publisher in separate files accompanying the journal article.

Abstract

We have developed and tested an evidence-based methodology for early stage identification of scientific discoveries. Scholarly publications are analysed to track and trace breakthrough processes, and their impact on world science. The focus in this study is on the incremental discovery of the Ubiquitin-mediated proteolytic system in the late 1970s by a small international team of collaborating researchers. Analysis of their ground-breaking research papers, all produced within in a relatively short period of time, and the network of citing papers shows the cumulative effects of the intense collaboration within a small group of researchers working on the same subject. Using the bibliographic data from the Web of Science database and the PATSTAT patents database in combination with expert opinions shows that these discoveries accumulated into a new technology. These first findings suggest that potential breakthrough discoveries can be identified at a relatively early stage by careful analysis of publication and citation patterns.

4.1 Introduction

4.1.1 The context of scientific breakthroughs

Can scientific breakthroughs be identified within a few years after the fact? What kind of sudden changes and distinctive structural developments do these events leave behind in the research literature? We study the observable effects of these breakthroughs on the research community. Our bibliometric approach relies on systematic large-scale searches within the worldwide scholarly and patent literature. In most cases, the precise moment in time when a breakthrough occurred cannot be precisely pinpointed – not even in retrospect. External observers and analysts, i.e., researchers who are not experts in the field under study, have to resort to indirect evidence and proxies. Assembling information and time-series data from bibliographic databases such as the Web of Science⁸¹ or SCOPUS⁸² enables non-expert analysts to identify significant short-term changes in the general pattern of knowledge creation processes and communication patterns within science within 2 years after publication of key research papers. By adopting this method, our early-detection approach differs from work carried out by others; Arbesman (2010), for instance, focuses on the ease with which scientific discoveries can be made; a Delphi approach applied to Japanese-German cooperation is presented in (Breiner et al., 1994); Chen et al. (2009) model transformative discoveries focussing on connections across structural holes in network representations of scholarly knowledge in scientific discovery; the focus of Julius et al. (1977) is on using expert knowledge to judge at an early stage whether a discovery is expected to be of importance for new technology that will effect R&D in the pharmaceutical sector; Leydesdorff and Rafols (2011) use geographical and cognitive diffusion to analyse emerging topics; Martin (1995) introduces a model of a foresight process to help identify scientific developments of strategic importance; Small (1977) answers the question: “Does the citation picture of a ‘specialty’ development match the perception of the specialists?”; Ponomarev et al. (2014b) use forecasting models to predict future citation patterns of individual papers to see if these papers are expected to exceed a high citation threshold after five years; in Ponomarev et al. (2014a) the usability of interdisciplinary in the subject categories, and geographical diversity is studied to see if these measures can be used to identify potential breakthrough papers.

Discoveries, breakthroughs and subsequent shifts in scientific ‘paradigms’ are characteristic for revolutionary science according to Kuhn, who distinguishes evolutionary ‘normal’ science and ‘revolutionary’ science (Kuhn, 1962). These structural shifts are marked by the gradual introduction and application of new theories, insights and concepts and their increasing dominance, and thus structurally changing the ‘cognitive’ fabric of a research domain (Andersen et al., 2006). However, science does not always progress in such a neat

⁸¹Web of Science is a product of Thomson Reuters

⁸²SCOPUS is a product of Elsevier B.V.

linear and cumulative way. Changes in theoretical frameworks proliferate slowly as the relevant scientific research community needs to be convinced of the need to alter its views and approaches. Radically novel approaches, new information and breakthrough finds, which are incompatible with the current dominant theoretical framework and beliefs within a research field, may suddenly appear on the scene and revolutionize the cognitive structure of this field of research within a brief span of time. Wray (2011, p.202) states: "According to Kuhn's mature view, a new theory is developed in a field in an effort to account for an anomaly that the accepted theory was unfit to account for." Wang et al. (2013, p.127) conclude: "Paradigm-changing discoveries have notoriously limited impact."

There is no generally accepted description, let alone a universal definition, of the term breakthrough that can count on full support throughout the international scientific community. We rely on Hollingsworth (2008, p.317) who states: 'A major breakthrough or discovery is a finding or process, often preceded by numerous small advances, which leads to a new way of thinking about a problem.' Building on this definition, we adopt Koshland's 'Cha-Cha-Cha' theory of scientific discovery, which distinguishes three general types of scientific discoveries (Koshland, 2007, p.761). He argues: "In looking back on centuries of scientific discoveries, however, a pattern emerges which suggests that they fall into three categories — Charge, Challenge, and Chance —". Where charge discoveries can be considered 'normal' science in the Kuhnian sense; the challenge discoveries and chance discoveries are examples of 'revolutionary' science marked by sudden changes in knowledge creation processes. Science progresses by building on the achievements of other scholars, as is expressed by the metaphor "If I have seen further it is by standing on the shoulders of giants".⁸³ But does this general metaphor apply to scientific discoveries of the Challenge type? According to Koshland (2007, p.761) 'Challenge' discoveries are a response to an accumulation of facts or concepts that are unexplained by or incongruous with the scientific theories of the time'. We use the term 'theory-changing breakthroughs' for discoveries of this type as they lead to alterations in theoretical concepts.

In some cases, the challenge lies within a small closely-knit, but geographically dispersed, group of researchers facing the same puzzling experimental results. As the research within this team progresses, their publications will cite each other's work. Several incremental 'minor' discoveries are made and published. The accumulation of their efforts, the teamwork discovery, will only be recognised and publicized by these 'insiders'. Their resulting (joint) research publications will therefore at first receive only a modest number of 'within-team' citations. The further diffusion of the ensuing new theoretical insights and the proliferation of their applications will result in a growing citation network. We hypothesize that citation patterns of such 'revolutionary' teamwork-based challenge discoveries differ from 'evolutionary' breakthroughs that are congruent with existing theories. We expect a socio-

⁸³This metaphor is ascribed to Isaac Newton but has to be attributed to Bernard of Chartres and was first recorded in the twelfth century (Merton, 1965, p.37)

cognitively dense but low-level citation network within the first couple of years after the discovery, followed by a sudden increase in citation numbers across the wider scientific community. This assumption is put to the test in our case study of a Nobel Prize winning discovery, which is introduced in the next section.

Several existing studies focus on research teams, and their role in the production of new knowledge. Weinberg (1970) focuses on the role of big team science; the focus in (Guimerà et al., 2005) is on mechanisms of team assembly. International scientific collaboration is discussed in (Luukkonen et al., 1992); Jones et al. (2008) focus on multi-university research teams. Wuchty et al. (2007) discuss the growing dominance of scientific research teams in knowledge production. (Darden and Maull, 1977) focuses on the role that ‘interfield theories’ play in the production of new knowledge; (Upham et al., 2010) discuss the advantages and disadvantages of ‘schools of thought’ in relation to the creation of new knowledge; the relation between collaboration and creativity is analysed in (Uzzi and Spiro, 2005); in Uzzi et al. (2013) the role of ‘boundary spanning ideas’ and especially ‘atypical combinations’ are discussed. In this study we focus on the question: “Is there a cognitive reason for this, as we know now, fruitful collaboration?”

4.1.2 Discovery of the Ubiquitin-mediated proteolytic system

The judgement of scientific experts in the specific field and on argumentation provided by other knowledgeable reviewers, in this case the Nobel Prize committee may be relied upon as a means of judging and contextualising the importance of a scientific discovery. In 2004, the Nobel Prize for Chemistry was awarded to Aaron Ciechanover, Avram Hershko (both scientists at the Technion, the Israel Institute of Technology at Haifa), and Irwin Rose (Fox Chase Cancer Center, Philadelphia, US) for the discovery of the Ubiquitin-mediated proteolytic system, which is crucial for the proper functioning of Eukaryotic⁸⁴ cells. Nath and Shadan (2009, p.421): Illustrate the function and importance of the Ubiquitin-mediated proteolytic system in the following way:

“The destruction of proteins is as important as their synthesis for the maintenance of protein homeostasis in cells. In eukaryotes, the ubiquitin-proteasome system is responsible for most of this protein degradation: the small protein ubiquitin acts as a death warrant, tagging and targeting other proteins to the large proteolytic chamber of the proteasome. ... It is now known that ubiquitin-mediated destruction plays a crucial part in cell-cycle regulation, DNA repair, cell growth and immune function, as well as in hormone-mediated signalling in plants. ... Given the central role of the ubiquitin system in diverse cellular processes, it is not surprising that its dys-

⁸⁴Eukaryotes have cells in which the genetic material is DNA in the form of chromosomes and is contained within a distinct nucleus. Eukaryotes include all living organisms other than the eubacteria and archaeobacteria.

function contributes to cancer and to neurodegenerative and immunological disorders. An understanding of the ubiquitin system is therefore important in devising treatments for such diseases ...

According to the Nobel Prize Committee (2004, p.3) *“The breakthrough came in 1980. It was described in two papers that were both communicated on 10 December 1979 to the journal Proceedings of the National Academy of Sciences of the USA.”* It continues with *“The unravelling of the Ubiquitin proteolytic system is not an exception to the rule that scientific discoveries are based on findings of others and that it can take a long period between the first preliminary findings and the breakthrough discovery.”* Ciechanover (2009, p.1) concurred with Koshland’s view of Challenge discoveries: *“... Could the Ubiquitin system have been discovered earlier? Possibly yes. ... As we now know, the system was not discovered by chance but rather by challenge — mostly as a natural response to developments in the field ... A new system and concept(s) were needed to explain all of these new findings and assumptions, gathering them under a unifying umbrella ...”*

4.2 Hypothesis, research questions, data and method

4.2.1 Hypothesis and research questions

The main hypothesis we want to test using the discovery of the Ubiquitin system as a test bed is ‘the observable changes in the socio-cognitive structure, that are visible as changes in the citation network of relevant bibliographic information, enable the detection at early stage of a scientific discovery that results in an identifiable shift in theoretical concepts.’ This hypothesis leads to the following research questions:

- How does the metaphor ‘science progresses by standing on the shoulders of others’ manifest itself in this discovery?
- What changes in the ‘socio-cognitive’ structure are visible in the period around a theory-changing breakthrough?
- Does the citation network of relevant publications support the idea of a ‘challenge’ breakthrough by showing only modest numbers of citations for the three landmark publications?
- Is it possible to perceive a gradual increase in the multidisciplinary of the citing publications?
- Can publications that are not directly manifest but that play an important role in the developments following the discovery be identified in the citation network around the landmark publications?
- Does the citation network show a ‘clustering’ of publications from a small group of scientists over a short period?
- Does the constraint we impose in using brief windows of time prevent the identification at an early stage of a theory-changing breakthrough?

4.2.2 Data

As data sources we used the on-line version of Thomson-Reuters Web of Science database (WoS) as well as CWTS's in-house version for the scholarly publications, and EPO's Worldwide Patents Statistics database (PATSTAT), the October 2012 edition, for the patent publications. All documents covered by the WoS up to the first quarter of 2013 and all patent documents published up to August 2012 were included in our search for relevant publications. In addition, we referred to the original 'paper' publications for those publications not covered by the WoS to reveal the citation information needed in our research.

Applying 'ubiquitin' as search topic to the WoS resulted in 28,778 relevant research publications of the document types 'article', 'letter' and 'proceedings paper' spanning the period 1975–2013. Searching PATSTAT for patent publications related to 'ubiquitin' resulted in 1,522 documents belonging to 682 distinct DOCDB⁸⁵ patent families that were applied for in the period 1985–2012.

4.2.3 Methods

We start with an analysis of the evolution of the research field by looking at the trends in the numbers of scholarly and patent publications. CWTS uses 251 distinct subject categories based on the subject categories Thomson-Reuters uses in the WoS database. Reclassification of all documents is carried out to reflect the most actual version of the classification scheme. In this context, a group documents is called 'mono-disciplinary' if all documents belong to one and the same subject category, and 'multidisciplinary' if multiple subject categories are assigned to the publications in the group. We look at the evolution of the multidisciplinaryity of the publications in this research field by analysing the evolution of the number of distinct subject categories assigned to the publications, and we conclude with the analysis of the citation networks centred on the breakthrough papers.

4.3 Results

4.3.1 Publication trends

Figure 4.1 shows the trend for scholarly publications as well as for patent publications. Several significant and visible changes in the first few years can be seen after 1980, when the breakthrough papers were published. A significant increase is clearly visible starting in 1983. Combining both figures, we see that this increase marks a pivotal moment, after which a continuous increasing number of scholarly publications can be found. Following this

⁸⁵DOCDB patent families are sets of equivalent patent publications referring to the same patent invention.

4.3 Results

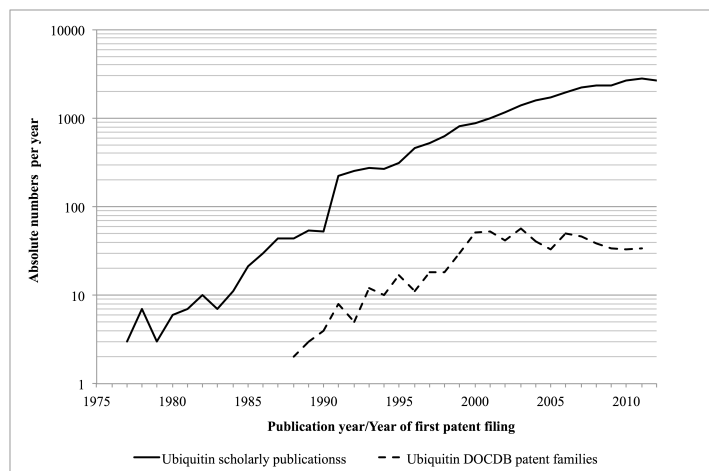


Figure 4.1: Numbers of scholarly publications (articles, letters and proceedings papers), patent applications (DOCDB patent families) per year (1975–2012)

pivotal moment in 1983, we notice the appearance of patent publication from 1988 onwards, indicating the gradual evolution of the scientific ideas into technology. The first patent publication found in this study was applied for in 1985. The long-term trend for Ubiquitin-related scholarly publications shows an exponential growth rate until it starts to deviate around 2003.

The trends shown in Figure 4.1 also illustrate that the Ubiquitin field matured into a substantial research area around 13 years after the main discovery. This moment is marked by the steep rise from 1990 to 1991. The number of patent families shows a significant and large increase in the period 1998–2000. This increase gives the impression of a switch to a new or improved technology, resulting in a higher level of patent applications.

4.3.2 Multidisciplinarity of the research field

As a measure of multidisciplinarity, we counted the number of different CWTs subject categories that correspond to the WoS-indexed publication output per publication year. As Ubiquitin-related research becomes more dispersed across related science fields, the number of publications in journals assigned to more and different subject categories will increase. This process reflects the pervasive spread of a new paradigm across the sciences, and the wider use of the knowledge on the Ubiquitin system. Figure 4.2 shows an overall fairly linear increase in the number of CWTs subject categories, totalling more than 100 in recent years. Deviations from this linear trend are visible at a more detailed level. From 1980 until 1985, the number of subject categories stayed more or less at the same level. In 1987 a jump took place and the number reached a plateau and stayed at this level until the steep rise in 1991. From

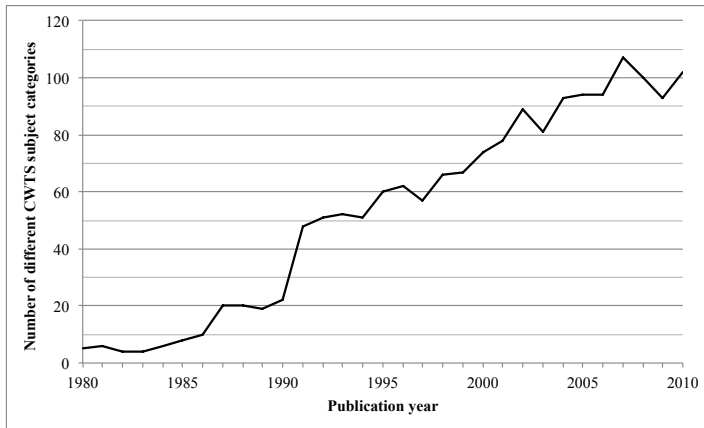


Figure 4.2: Number of different CWTs subject categories assigned within a year to Ubiquitin-related publications (1980-2010)

1991 onwards a more or less linear growth is visible.

4.3.3 Citation network

Network analysis

Aaron Ciechanover (2009, p.2) mentions "... but it is nevertheless possible to point to a single article, or a handful of articles, that laid the foundation to the discovery. This is the case with the Ubiquitin system, where the pioneering findings were described in a single article published in this journal ...". The publication referred to is (Ciechanover et al., 1978) that was published almost two years before the two publications (Ciechanover et al., 1980b) and (Hershko et al., 1980). According to the Nobel Prize Committee, these latter two publications together describe the paradigm-involving breakthrough. We refer to these three publications as the 'landmark publications'.

To analyse the role the breakthrough papers played in the development of the 'Ubiquitin' field, we constructed an extended citation network consisting of publications directly citing the 'landmark publications' and of publications cited by these three publications. In constructing this network, only publications containing original research are included; reviews, editorials, corrections and the like were therefore omitted. The citation network contains 174 publications and spans the period 1951-1986. To find the publications that, based on citation relations, were the most influential, we included all 1,023-citation relations between documents within this network. For every node in this network the 'indegree' was calculated and used as measure of importance of the node. Indegree indicates the number of times a publication is cited by publications within this time-constrained network, and it is therefore less than the overall number of times a publication is cited according to the Web of Science database.

4.3 Results

Table 4.1: Key figures for the citation network for the period 1951–1986 with (Ciechanover et al., 1978), (Ciechanover et al., 1980b), and (Hershko et al., 1980) as focus points

Number of nodes	174
Number of vertices (links) within this network between the nodes	1,023
Publication	Indegree
Hershko et al. (1980)	83
Ciechanover et al. (1980b)	53
Wilkinson et al. (1980)	42
Ciechanover et al. (1978)	39
Ciechanover et al. (1980a)	33

The publications are divided into two groups. One group consists of publications written by authors and co-authors of the Technion Institute and Fox Chase Cancer Center (Technion-Fox Chase group) and a second group consists of publications by authors not in the Technion-Fox Chase group. We present key figures for this network in Table 4.1. The indegree divides the publications into two distinct groups. There are five publications in this network, all from the Technion-Fox Chase group, with an indegree above 30. All other publications have an indegree of 15 or less; meaning that these publications play a less prominent role in Ubiquitin research. The distribution of the publications within this network with respect to their indegrees is given in Figure 4.3. The five publications with the highest indegree stand out noticeably and are clearly marked. Two of these five publications, (Ciechanover et al., 1980a) and (Wilkinson et al., 1980) became visible as concentration points in the extended citation network as they were published shortly after the Nobel Prize-winning publications.

Citation network of the Ubiquitin discovery

To investigate the direct and indirect cognitive relations between these three publications, we created a citation network for the discovery of Ubiquitin (Figure 4.4). The network is based on the citation links for the ‘landmark publications’. The opaque disks present these three publications. Open circles represent publications with an author or co-author from the Technion Institute. Non-Technion publications are represented by the name of the first author and the year of publication. The publications referenced by these three publications are also included. Continuous lines indicate citation relations between documents authored or co-authored by researchers from the Technion Institute.

Dashed lines in this figure indicate references to non-Technion publications. The citation relations indicate that the research was directly based on scientific knowledge published from 1951 onwards. The publication in 1951

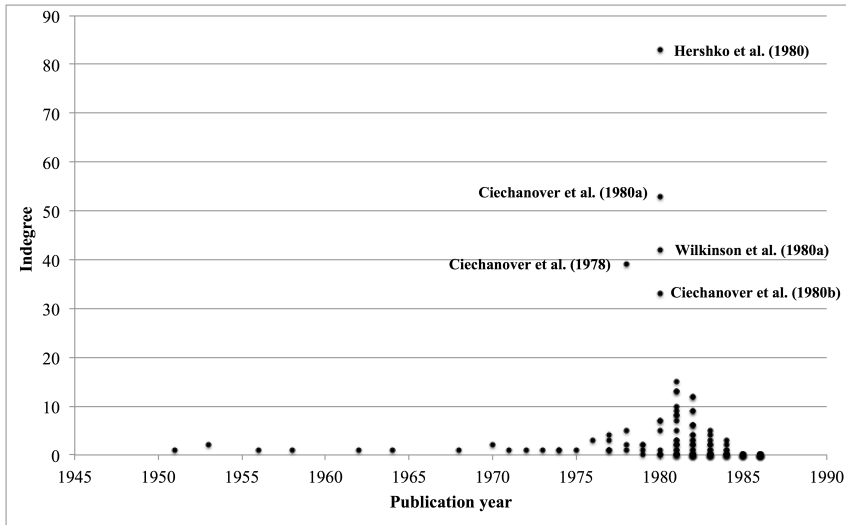


Figure 4.3: Indegree for all publications within the citation network (1951–1986). The dots represent individual publications. The labelled dots represent the five publications with an indegree above 30

is Lowry et al. (1951), which is a well-known example of a very highly cited (> 300,000 citations) publication (Moed, 2005, p.82).

Figure 4.5 shows an annotated detail of Figure 4.4, and is based on the publications from the two institutes up to the discovery publications. Authors and co-authors of the publications, their affiliations, and citation relations are shown. The octagons represent affiliations; ellipses represent authors, and disks publications.

The affiliations and the authors are placed on the left-hand side (organisational structure); the right-hand side (time-dependent citation structure) contains the publications with the time dimension on the horizontal axis. The opaque ellipsis represents the three Nobel Prize Laureates; the opaque disks represent the 'landmark publications'. These two figures clearly show that the Technion Institute is at the centre of the discovery of the Ubiquitin proteolytic system. The scientific knowledge in the 'landmark publications' comes from publications that are closer on a time line to each other when approaching 1980; the citation network becomes denser in the time dimension.

The socio-cognitive bridge: the Technion – Fox Chase research team

Ciechanover (2014): “The cooperation between the Technion Institute and the Fox Chase Cancer Center was sought by the researchers at the Technion Institute to solve the problem of the involvement of ATP⁸⁶ (energy) in the

⁸⁶Adenosine triphosphate (ATP) is a molecule that acts as energy source within cells (footnote added by the authors)

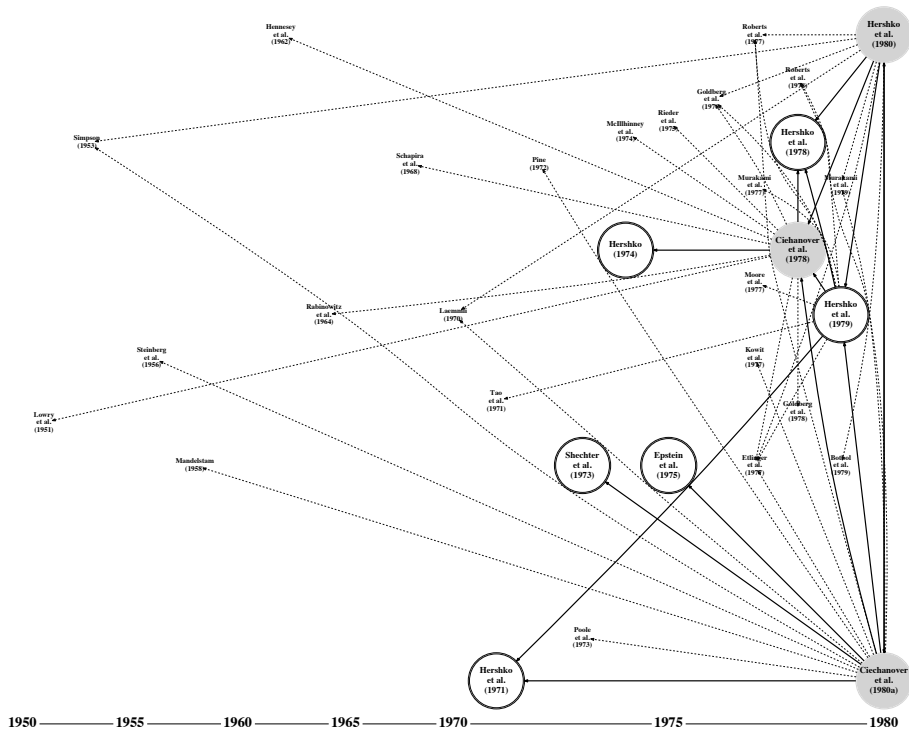


Figure 4.4: Citation network of the Ubiquitin discovery based on publications cited by the 'landmark publications'

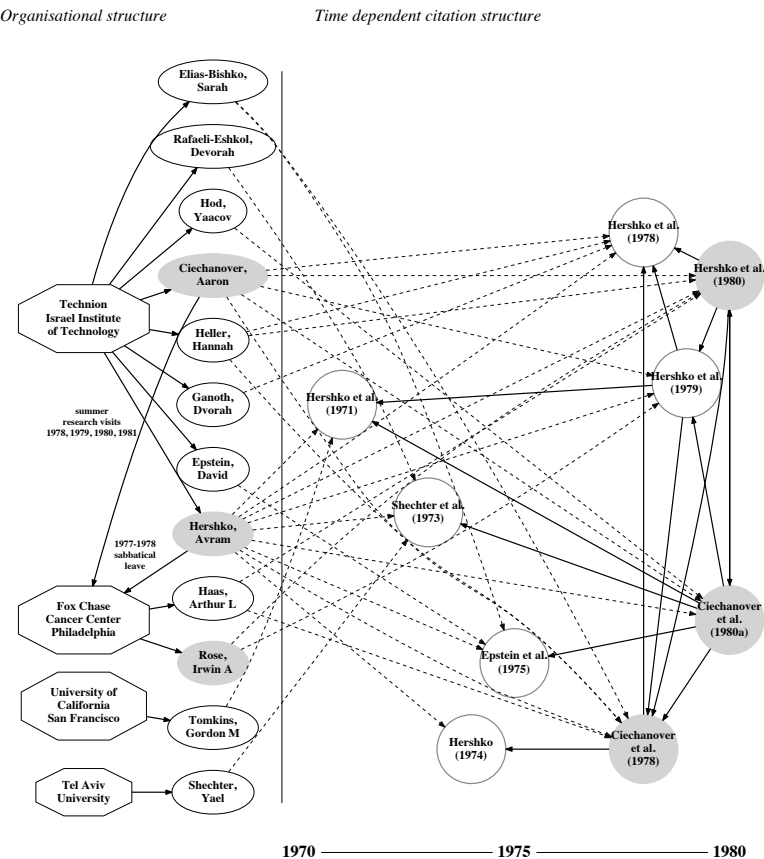


Figure 4.5: Organisational-temporal anatomy of the discovery of Ubiquitin. Shown are publications, (co) authors, research institutions, and interrelations. The opaque (grey) ellipses present the Nobel Prize laureates, and the opaque (grey) disks the 'landmark publications'. Solid arrows represent affiliation-author relations and citation relations; (co) author-publication relations are represented by dashed arrows

4.3 Results

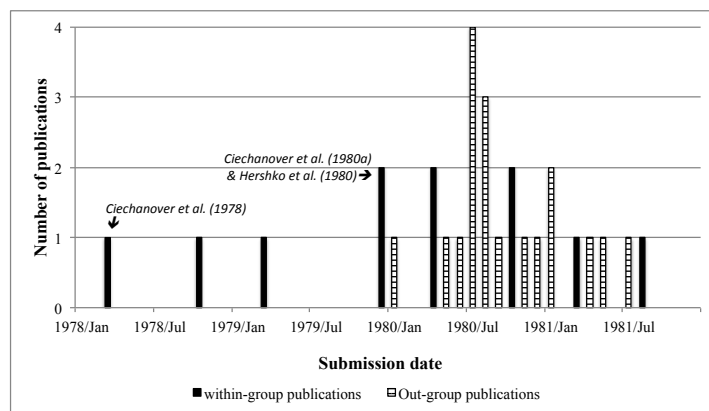


Figure 4.6: Within-group and out-group publications connected by citation relations with the ‘landmark publications’ (1978–1981)

ubiquitin process; Dr. Rose’s group offered fundamental help in solving this issue.”. Ciechanover, Herskho, Rose, Haas, and Wilkinson formed the core of this research team. The first publication related to Ubiquitin in which at least one author of the core research team from each institute is involved is (Rose et al., 1979); the last publication is (Hershko et al., 1984). As the team members share information it is expected that they cite their publications before other researchers do. Figure 4.6 shows for the period 1979–1981 publications with citation relations with the ‘landmark publications’; publications from authors within the core-group (within-group publications) and those from authors outside (out-group publications) are clearly marked. The date a manuscript was submitted to the publisher is used to place the publications on a time line.

Knowledge profiles based on WoS subject categories assigned to the publications of the Nobel Prize laureates before the collaboration was established are presented in Figure 4.7; overlap and complementarity in subject categories is clearly visible.

4.3.4 Citation characteristics of the ‘landmark publications’

Which publication is considered a highly cited paper?

How do the documents presented compare when taking the number of times they are cited as a measure? Table 4.3 gives an overview. As the table shows the core documents in the Ubiquitin case show modest citation behaviour. Redner (2005, p.25) defines, arbitrarily he admits, a ‘discovery paper’ as having at least 500 citations. Using this criterion, none of the Ubiquitin-related papers can be considered a ‘discovery paper’. The WoS subject categories ‘Biochemistry & Molecular Biophysics’ and ‘Biology’ are assigned to the four publications (Ciechanover et al., 1980a,b; Hershko et al., 1980; Wilkinson et al.,

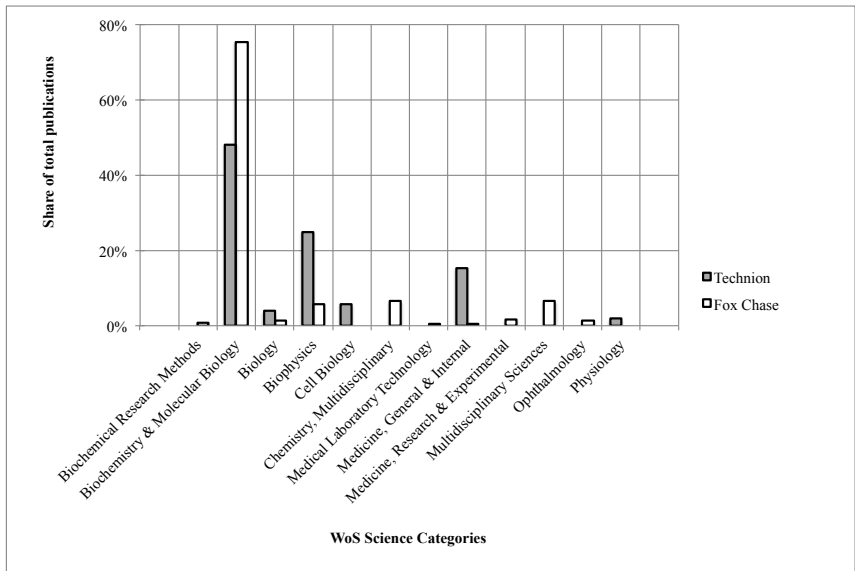


Figure 4.7: Knowledge profiles of the Technion Institute (Ciechanover and Hershko) and Fox Chase Cancer Centre (Rose) based on the WoS subject categories of publications up to 1978 covered in the Web of Science database

1980). Citation distributions were calculated for all publications of types ‘article’ and ‘letter’ published in 1980 to which these subject categories were assigned; the four publications were excluded from the calculations. We present the results in Table 4.2. All four publications are among the highest cited publications published in 1980 for the two WoS Subject categories.

The role of (Ciechanover et al., 1978)

The citation history for (Ciechanover et al., 1978) started in 1979 with a number of group self-citations by authors and co-authors from the Technion Institute. It seems that the real value of the discovery presented in this publication was not recognised at this time outside the inner circle of researchers working on the same subject. (Ciechanover et al., 1978) was published in *Biochemical and Biophysical Research Communications (BBRC)*. Compared with all publications published in 1978 in this journal, it has proven to be one of the most cited publications. The distribution and the relative position of the paper within this distribution are presented in Figure 4.8. It has received fewer citations than (Ciechanover et al., 1980b) and (Hershko et al., 1980). These latter two publications were published in *the Proceedings of the National Academy of Sciences (PNAS)*, and are among the top 6% most cited PNAS-publications published in 1980.

4.3 Results

Table 4.2: Key citation figures for the ‘core’ document shown in Figure 4.5. Hershko (1974); Hershko et al. (1978) are conference papers for which no citation information is available.

Publication	Cumulative number of citations			
	≤12 months	≤24 months	≤36 months	Until September 2013
Hershko and Tomkins (1971)	7	25	31	294
Shechter et al. (1973)	1	4	6	17
Hershko (1974)	.	.	.	.
Epstein et al. (1975)	4	15	22	116
Ciechanover et al. (1978)	0	4	16	283
Hershko et al. (1978)	.	.	.	.
Hershko et al. (1979)	6	18	30	137
Ciechanover et al. (1980b)	4	22	42	337
Ciechanover et al. (1980a)	1	16	31	159
Hershko et al. (1980)	7	30	59	467
Wilkinson et al. (1980)	3	26	51	250

Table 4.3: Based on the number of times cited the position of the four ‘core’ publications from 1980 compared to the publications in the ‘reference set’ is shown.

Publication	≤ 2 years	Up to 2013
Ciechanover et al. (1980b)	Top 2%	Top 1%
Ciechanover et al. (1980a)	Top 4%	Top 5%
Hershko et al. (1980)	Top 1%	Top 1%
Wilkinson et al. (1980)	Top 1%	Top 2%

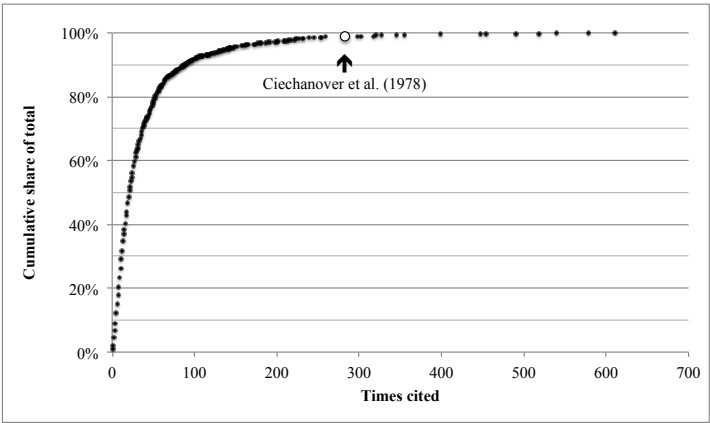


Figure 4.8: Cumulative distribution of the number of publications in Biochemical and Biophysical Research Communications published in 1978 versus the number of times cited (1978–2013)

4.4 Discussion

The aggregate-level trend data depicted in Figure 4.1 show several development stages in the number of scholarly publications, notably in 1984–1985, 1990–1991, and 2008–2009. These upswings in the time series suggest that prior scientific discoveries caused sudden increases in research activity and a concomitant rise in the number of publications. The type of the discovery and its impact on R&D processes worldwide cannot, however, be determined from these trend data.

For the patent data, we notice that the applications start to appear in 1985 after the rise in the number of scholarly publications in 1984–1985. This rise also coincides with the sudden increase in the number of CWTs subject categories (Figure 4.2). This indicates that the scientific knowledge about Ubiquitin became mature enough to be used in patent applications. The level switch in the number of patent application that we notice in 1999–2000 is an indication that new or improved usage of the knowledge was found to be the basis for further technological developments.

In Figure 4.1, which focuses on the period when the breakthrough discovery was made, a steep rise is noticeable in the period 1983–1987. The evolution of the number of patent applications, the time delay with the upswing of scholarly publications, and the simultaneous level change in multidisciplinary (Figure 4.2) indicate that the scientific knowledge became sufficiently well developed to be transformed into patentable technological applications (inventions).

The analysis of the citation network (Figure 4.5) shows that researchers from the Technion-Fox Chase group played a central role in Ubiquitin development. The discovery of the Ubiquitin-mediated proteolytic system was at first not recognised outside the inner circle of the research group around Ciechanover and Hershko. Hershko spent a sabbatical leave at the Fox Chase Cancer Center in Philadelphia in 1977–1978; Ciechanover visited this institute in the summers of 1978–1981 (Nobel Prize Committee, 2004). These stays and visits, and also the publications by members of the Technion - Fox Chase research team signal the importance of this collaboration for Ubiquitin research.

Figure 4.4 and Figure 4.5 show that the three ‘landmark publications’ are related by citation links, and that they also share a group of authors and co-authors, of which most are affiliated to the Technion-Fox Chase group. These authors form a research group based in two institutes that originated at the Technion Institute. Various aspects of scientific research groups are discussed in Guimerà et al. (2005); Jones et al. (2008); Stokols et al. (2008); Uzzi et al. (2013); Uzzi and Spiro (2005); Weinberg (1970); Wuchty et al. (2007).

Of the nine publications of interest, five were published in the period 1978–1980 (Table 4.1). Within the triadic network formed by (Ciechanover et al., 1980b, 1978; Hershko et al., 1980), a fourth publication (Hershko et al., 1979) seems to act as an information bridge between these three publications.

In the long run, all the three landmark publications receive a respectable,

but not excessive, number of citations. (Ciechanover et al., 1978) starts receiving a modest number of citations in the second year after publication. From year two on, the number of citations to (Ciechanover et al., 1980b; Hershko et al., 1980) are relatively high but not very extreme. Based on the citations received, the 1978 paper at first went almost unnoticed compared to the two 1980 papers. This is especially true if the paper is compared with another publication (Hershko et al., 1979) that received much more attention. There are several plausible explanations for this behaviour, for instance (1) the research field is becoming more mature and starting to realise the importance of the discovery, and (2) the use of different journals *Biochemical and Biophysical Research Communications* (Ciechanover et al., 1978) vs. *Proceedings of the National Academy of Sciences* (Ciechanover et al., 1980b; Hershko et al., 1980, 1979) with, partially, different audiences, to name just two of the possibilities.

The citation network for all publications directly cited by the three 'landmark publications' or citing them in the period 1951–1986 shows two publications, (Wilkinson et al., 1980) and (Ciechanover et al., 1980a), that also seem to play an important role. The role of the findings presented in these two publications is mentioned in (Hershko and Ciechanover, 1982, p.350). These two publications describe in combination the definitive identification of the component known as APF-1 as Ubiquitin.

The knowledge profiles (Figure 4.7) largely overlap but show some differences. The Technion team seems to be more focused on medicine, biophysics, and physiology. The conclusion arises that a cognitive bridge was created between two groups with at least partial complementary knowledge and that this combination facilitated the disentangling of the nature of the Ubiquitin system. This makes it plausible that the collaboration between the two groups of authors introduced new combined theoretical insights that played a significant role in the further development of the original discovery; Ciechanover (2014) supports this view.

Other information sources and related bibliographic data have to be examined to determine the nature and background of the breakthrough, in order to differentiate the roles the various key publications played. The discovery is a product in a wider field of research on protein degradation. We note that the discovery of the Ubiquitin-mediated proteolytic system emerged from research executed by a small group of interacting scientists, as far as we can tell from bibliographic information. From Figures 4.4 and 4.5 we conclude that the closer we approach the year 1980, the denser the network of publications linked by citation relations becomes.

The evolution of networks of researchers and institutions becoming active in this research field can be taken to reflect the diffusion of the new theoretical concepts within the scientific community. In this study our goal was to answer the following research questions:

How does the metaphor 'Science progresses by standing on the shoulders of others' manifest itself in this discovery?

Our research shows that in the period around the breakthrough discovery a small group of intensively co-operating researchers were endeavouring to attain the breakthrough discovery through bootstrapping. At a later stage, new researchers gradually enter the research field and ‘stand on the shoulders’ of the achievements of the group of original researchers.

Is there a cognitive reason for this, as we know now, fruitful collaboration?

According to Ciechanover (2014) the collaboration between the two research groups was instigated by the fact that the research group at the Fox Chase cancer centre was needed to help elucidating the role of ATP in the Ubiquitin-mediated proteolytic system.

What changes in the ‘socio-cognitive’ structure of the discovery of Ubiquitin are visible in the period around a theory-changing breakthrough?

The socio-cognitive structure of the discovery of Ubiquitin shows that the very small original research group enlarged as result of a sabbatical visit of Herskho, and several visits by Ciechanover. The start of these visits marks the moment publications from both groups become part of the same citation network that led to the landmark publications for which the Nobel Prize was awarded.

Does the citation network of relevant publications support the idea of a challenge breakthrough by showing only modest numbers of citations for the three landmark publications?

The ‘landmark publications’ are among the most cited publications in the field but the number of citations is modest; at most 30 within 36 months. The citation network shows several other publications with a comparable number of citations. The breakthrough discovery did not lead to an explosion of research activity, but proliferated in subsequent iterative advances.

Is a gradual increase in the multidisciplinary of the citing publications noticeable?

A linear increase in multidisciplinary is visible (Figure 4.2) over a long period. For the period 1980–1984, it remained stable at a modest level. Instances of abrupt increases that might be the result of ‘follow-up’ breakthroughs are also visible.

Can publications that are not directly manifest but that play an important role in the developments following the discovery be identified in the citation network around the landmark publications?

Analyses of the citation network up to 1983 revealed two publications, published in 1980 shortly after the two 1980 landmark publications, which also seem to play a central role in the further development of research on the Ubiquitin-mediated proteolytic system (Hershko et al., 1982)

Does the citation network show a ‘clustering’ of publications from a small group of scientists over a short period?

A ‘clustering’ can be perceived of publications from a small group of collaborating authors in a relative short period around the theory-changing breakthrough discovery (Figure 4.4 and Figure 4.5). Publications from this group of collaborating authors show a ‘within-group’ advantage (Figure 4.6).

Does the constraint we impose in using short time-windows prevent the identi-

fication at an early stage of a theory-changing breakthrough?

If we consider publications over the period of four years, 1978–1982, structural changes in the citation network around the breakthrough moment are manifest. The papers describing the ubiquitin-discovery received a modest number of citations. This modest number of citations is in line (Wang et al., 2013, p.127) ‘Paradigm-changing discoveries have notoriously limited impact’, if we interpret ‘impact’ as ‘citations received’.

4.5 Concluding remarks

This study finds that there is bibliometric evidence supporting the idea of a challenge breakthrough. The breakthrough discovery of the Ubiquitin-mediated proteolytic system is the result of several incremental discoveries distributed over a number of publications that are linked by citation relations. The individual publications are not exceptional in the number of citations they received, and are by a small group of collaborating scientists. Five publications seem to form the ‘heart’ of the discovery, and constitute the basis for further development. These publications stand out in indegree (i.e., number of citations received).

Another conclusion is that “standing on the shoulder of others” does not adequately describe the progress of science at the ‘micro’ level of this individual theory-changing discovery, i.e. challenge discovery, as the discovery is the result of the work of a small group of closely collaborating scientists. This group of scientists was in some respects, at least temporarily, cognitively isolated as result of reluctance in the scientific community to accept new theories. Wang et al. (2013) cite (Kuhn, 1962) as “...precisely because the more a discovery deviates from the current paradigm, the longer it takes to be appreciated by the community ...” Bibliographic information shows a within-group advantage (Figure 4.6), and furthermore that the scientific knowledge in the ‘landmark publications’ comes from publications that are closer on a time line to each other when approaching 1980.

Analysis of the publications and their citation relations in this study show effects on a time scale that might be appropriate for use as early stage (i.e., within 12 to 24 months) identification of a theory-changing breakthrough. It needs to be proven if our hypothesis

“observable changes in the socio-cognitive structure that are visible as changes in the citation network of relevant bibliographic information, enable the detection at early stage, within 24 months, of a potential scientific discovery that results in an identifiable shift in theoretical concepts”

holds in general. Redner (2005, p.52) defines a discovery paper as having at least 500 citations, Wang et al. (2013) point to the fact that publications with exceptional long-term impact appear to be the hardest to recognize on the basis of their early citation patterns, van Noorden et al. (2014) find that none

of the papers announcing some of worlds most famous discoveries come anywhere close to ranking among the 100 most highly cited publications of all time, and Ioannidis et al. (2014) find that authors conclude that their most cited papers are the most important and that the chart-topping work was evolutionary and not revolutionary. To alleviate this issue this study deviated from 'simple' citation analysis by incorporating in the analysis other information related to the publications. Careful analyses of citations at early stage revealed the inception of a core research group. The members of this research group were responsible for all citations during the first two years after publication of the discovery. From the moment the 'breakthrough' discovery is recognised by other scientists 'out-group citations' become more and more dominant.

A follow-up study should focus on operationalizing a method to identify from bibliographic data a 'core research group'. Furthermore this study should provide threshold values for the 'within-group citations' to 'out-group citations' ratio that would hint at a theory-changing discovery. An identified 'core research group' in combination with the value for the 'within-group citations' to 'out-group citations' ratio should help in generalising the results of this case study, and answer the question 'to what extent is the identification of potential theory-changing breakthroughs possible by searching for 'core research groups in early-stage citation networks?'

4.6 Acknowledgements

The authors sincerely thank professor Aaron Ciechanover for his willingness to answer questions posed and pointing to additional relevant literature. The authors also thank the reviewers for their comments and suggestions.

Appendix 4.A Supplementary material

4.A.1 Publications cited by and citing the 'landmark publications'

For the period 1950–1986 we provide a picture of the citation network showing all publications citing the 'landmark publications', and publications cited by the 'landmark publications'. For all publications the citation relations with all other documents within this set of publications are shown. (see Figure 4.11) on page 86. (Ciechanover et al., 1978) is represented in this picture by an orange disk. The two red disks represent (Ciechanover et al., 1980b; Hershko et al., 1980), and yellow disks represent (Ciechanover et al., 1980a; Wilkinson et al., 1980). All disks with a light blue colour represent publications by researchers at either the Technion Institute or the Fox Chase Cancer Centre. The disks are labelled with the first author, the publication year and the access number of the publication in the Web of Science database.

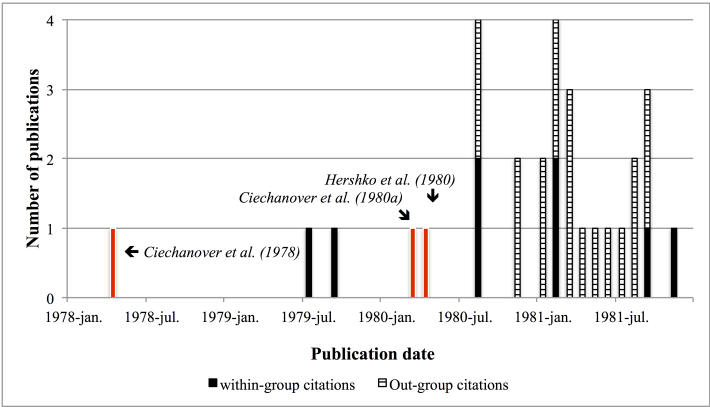


Figure 4.9: Within-group and out-group citations for the ‘landmark publications’ (1978–1981) (publication dates)

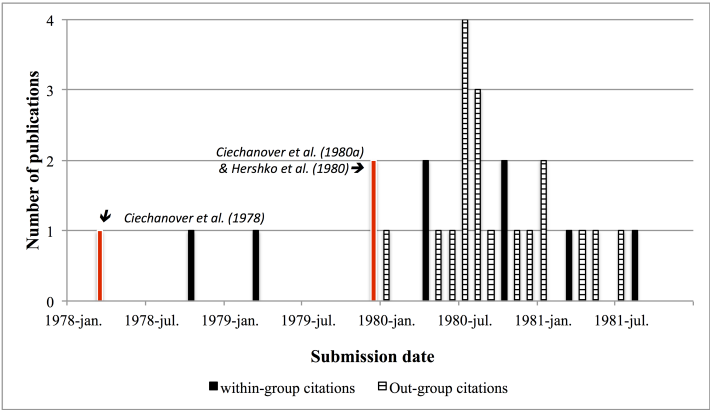


Figure 4.10: Within-group and out-group citations for the ‘landmark publications’ (1978–1981) (submission dates)

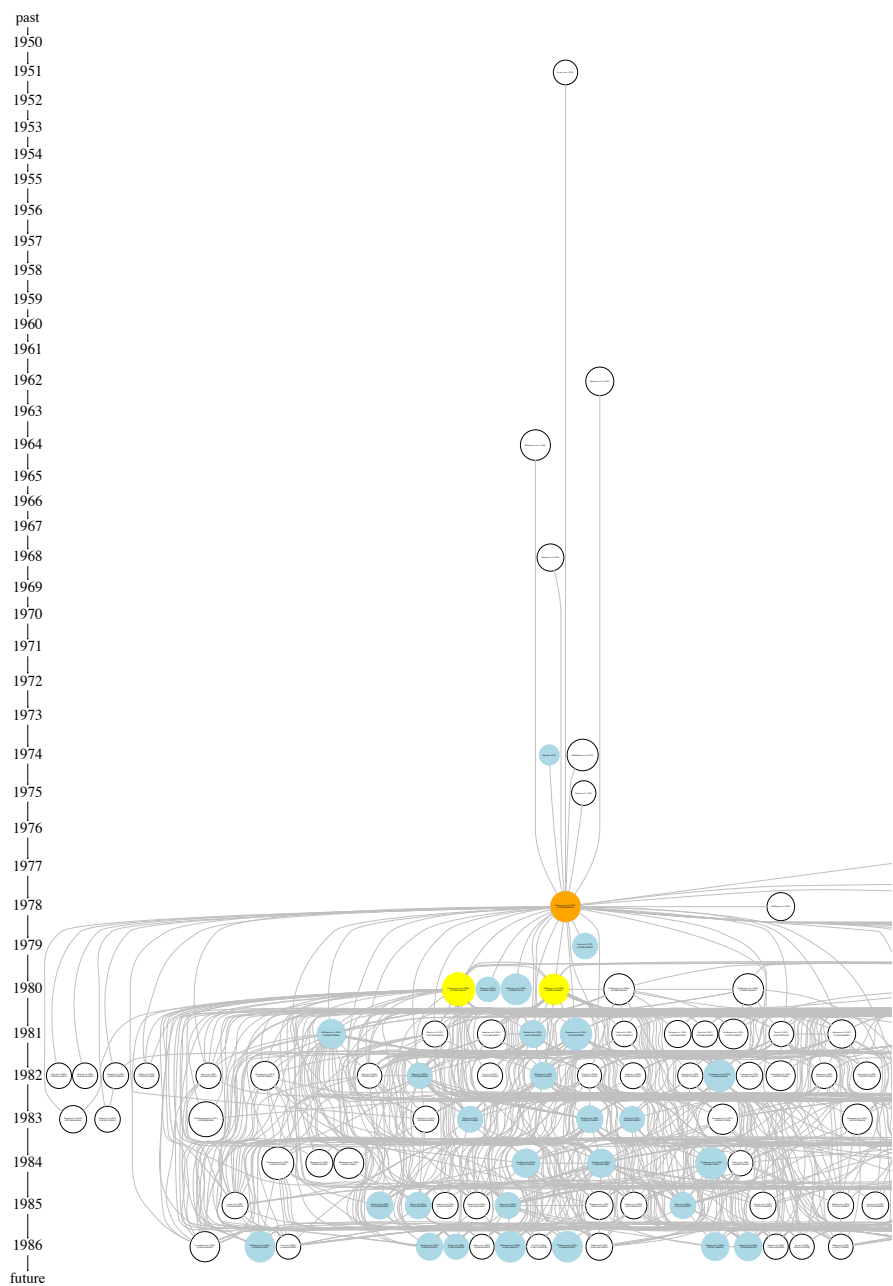
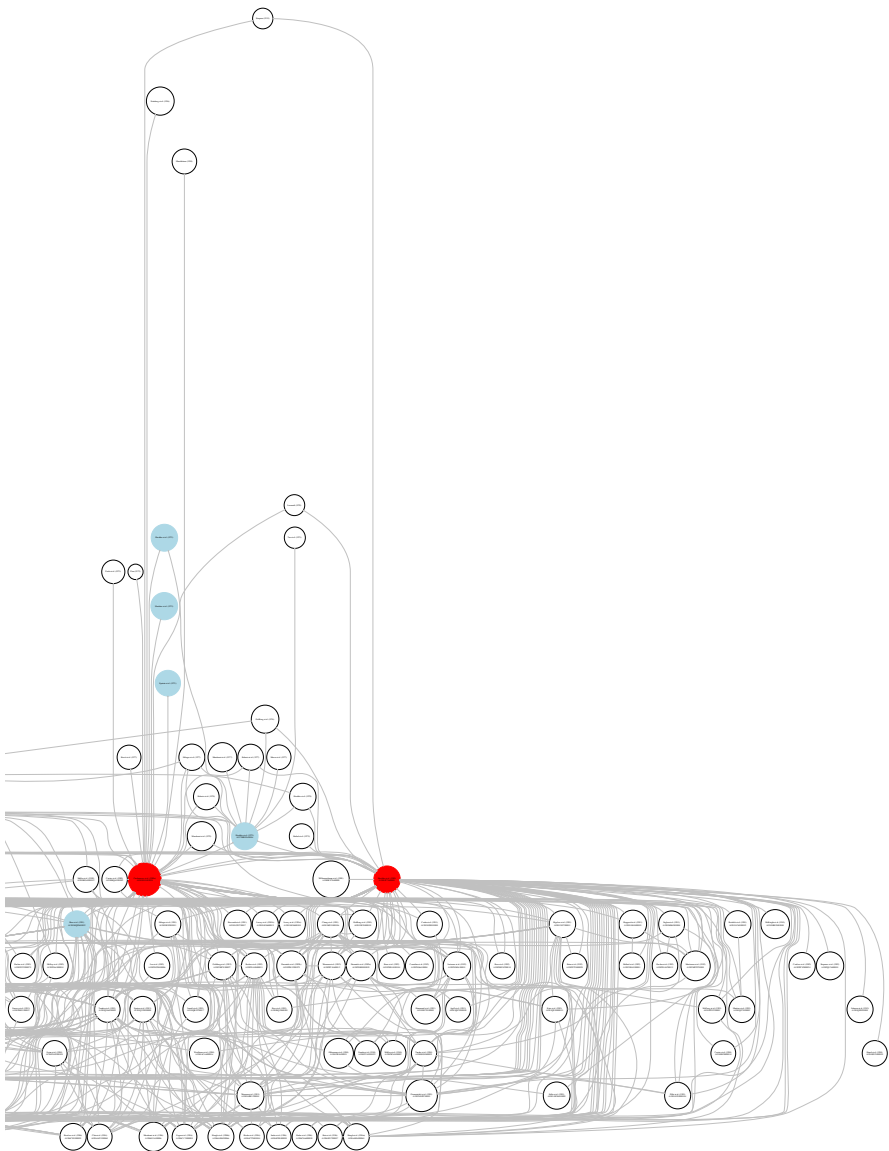


Figure 4.11: Publications cited by and citing the 'landmark publications' (1950–1986). In Section 4.A.1 on page 84 the meaning of the colors of the discs is explained.



4.A.2 Citation relations in the temporal dimension

At a 'microscopic' level we looked at the moment publications cite the 'landmark publications' to search for 'within-group' advantages. As a first approximation we used the publication dates as they appear in the Web of Science database. As is illustrated below the publication date is inappropriate to place publications on a time line in order to get a precise picture of the dynamic behaviour of research on a fine-grained time scale.

In Figure 4.9 the results obtained by using publication dates are shown for the publications citing the 'landmark publications' in the period 1978–1981. On the basis of this figure several observations can be made:

1. The relative long period before within-group citations to (Ciechanover et al., 1978) appear → this seems contradictory to the idea that the researchers in the research team found a major discovery;
2. It took more than 2 years after the publication of (Ciechanover et al., 1978) before the first citation from a researcher outside the Technion-Fox Chase team appeared in July 1980;
3. More striking is the fact that (Ciechanover et al., 1980b) is published one month before (Hershko et al., 1980) → this is suspicious as both publications cite each other, and are submitted to the same journal on the same date by the same corresponding author.

Based on these observations we collected the submission dates of the relevant manuscripts and used this information to place the publication on a time line. Figure 4.10 shows the results. The two publications published in the second half of 1979 that cite (Ciechanover et al., 1978) are now clearly separated in time and are closer in time with this publication.

The submission date, that is regrettably not available in the bibliographic databases, is a better approximation of the moment authors' think the findings of their research are mature enough to be published. This date is closer to the discovery than the publication date is. It is well known that several conditions influence the time lag between the submission of a manuscript and its publication.

On the dates at least two remarks can be made that we will not address any further in this study.

1. It is not certain that the submission date is the date (a version of) the manuscript was submitted to a publisher for the very first time; the manuscript might for instance be rejected on earlier occasions. The history of a manuscript before it is published is in general not known;
2. The publication dates suffer from the fact that journals have different publication frequencies. For instance in which month is a yearly, quarterly, or bi-monthly publication published? In our analysis publishing is used in this context as 'making known to the audience'. The publishing date should reflect the moment the manuscript becomes available to the audience.

Chapter 5

Case study #4: Introns

The content of this chapter is equivalent to the manuscript published as:

Winnink, J.J., Tijssen, Robert J.W., and van Raan, A.F.J. (2013). The discovery of 'introns'; an analysis of the science-technology interface. In Hinze, S. and Lottman, A., editors, *Translational twists and turns: science as a socio-economic endeavor - Proceedings of STI 2013 Berlin (18th International Conference on Science and Innovation Indicators)*, pages 427-438.

The only differences between the published article and the text in this chapter are of typographic nature to assure that the layout is in line with the typographic design of this PhD thesis.

Abstract

The study addresses visible clues and empirical data that can be used to discover early stage breakthroughs in science that evolve into new technological developments. We analyse bibliographical information from scholarly publications, patent publications, and links between patent and scholarly publications. Our overall goal is to develop an analytical methodology for pinpointing the stage in which fundamental discoveries occur. This particular case study focuses on the discovery of 'introns' in chromosomes. We discuss two breakthrough events related to this discovery, classify them according to the 'Cha-Cha-Cha' theory, and show that the second breakthrough reflects a 'phase shift' where the scientific discovery moves into the first stages of technological development.

5.1 Introduction

Our longitudinal bibliometric analyses of 'breakthrough processes' aims at obtaining better insight into general patterns that characterise 'revolutionary'

R&D dynamics (Winnink and Tijssen (2011); Winnink (2012)). We use the ‘Cha-Cha-Cha’ theory Koshland (2007) to classify breakthrough discoveries into distinctly different types. A short description of the ‘Cha-Cha-Cha’ theory is given in the next paragraph. The subject of this case study is the discovery of so-called ‘introns’⁸⁷. We categorise this discovery as a ‘Challenge’ because it can be characterised as posing a challenge that required a major ‘paradigm shifting’ adaptation of previous theories to explain new experimental observations (rather than solving an ‘obvious’ scientific problem).

The underpinning scientific discoveries were published in 1977 (see for instance Gelinas and Roberts (1977); Roberts (1993); Sharp (1993)) and revealed that chromosomes comprise of a mosaic of ‘exons’, used to form new copies of genes, and of ‘introns’ that appeared to be non-functional ‘interspaces’. These discoveries proved the assumption ‘chromosomes for prokaryotic⁸⁸ and eukaryotic⁸⁹ organisms are similar in structure’ to be wrong. The 1993 Nobel Prize in Physiology or Medicine was awarded to Philip A. Sharp and Richard J. Roberts for their pioneering work on ‘introns’. Sharp (1993) describes in his Nobel Prize lecture the ‘intron’ problem as follows:

“By the late 70s the physical structure of a gene was firmly established from work in bacteria. The sequences of the gene, the RNA and the protein were colinearly organized and expressed. Since the science of genetics suggested that genes in eukaryotic organisms behaved similarly to those of prokaryotic organisms, it was naturally assumed that this bacterial gene structure was universal. It followed that if the gene structure was the same, then the mechanisms of regulation were probably very similar, and thus what was true of a bacterium would be true of an elephant. However, many descriptive biochemical aspects of the genetic material and its expression in cells with nuclei suggested that the simple molecular biology of gene expression in bacteria might not be universal”

The two main research questions we address in this study relate to how these revolutionary breakthroughs can be identified and monitored over time:

- Can such a breakthrough, involving a paradigm shift, be identified and characterized in terms of bibliometric variables and indicators?
- Can one identify, bibliometrically, the stage in the development process where scientific knowledge is being used for work on science-based technologies?

⁸⁷The word intron is derived from ‘intragenic region’

⁸⁸Prokaryotic organisms are microscopic single-celled organisms, e.g. bacteria and cyanobacteria, that have neither a distinct nucleus with a membrane nor other specialized organelles.

⁸⁹Eukaryotic organisms are organisms consisting of a cell or cells in which the genetic material is DNA in the form of chromosomes contained within a distinct nucleus. Eukaryotes include all living organisms other than the eubacteria and archae.

5.2 Conceptual framework

Our analytical approach builds on ‘Cha-Cha-Cha’ theory by Koshland (2007), who’s work on discoveries

‘In looking back on centuries of scientific discoveries, however, a pattern emerges which suggests that they fall into three categories — Charge, Challenge, and Chance — that combine into a “Cha-Cha-Cha” Theory of Scientific Discovery.’,

resulted in the following general typology that differentiates discoveries based on their nature:

Charge discoveries solve problems that are quite obvious — cure heart disease, understand the movement of stars in the sky — but in which the way to solve the problem is not so clear.

Challenge discoveries are a response to an accumulation of facts or concepts that are unexplained by or incongruous with scientific theories of the time. The discoverer perceives that a new concept or a new theory is required to pull all the phenomena into one coherent whole.

Chance discoveries are those that are often called serendipitous⁹⁰ and which Louis Pasteur felt favoured “the prepared mind.” In this category are the instances of a chance event that the ready mind recognizes as important and then explains to other scientists.’

Charge discoveries can be considered ‘normal science’ in Kuhn’s (Kuhn, 1962) terminology. Challenge and Chance discoveries can be seen, within this particular context, as ‘revolutionary science’. Further work by Hollingsworth (2008) defines a ‘major breakthrough’ or a ‘discovery’ as: “a finding or process, often preceded by numerous small advances, which leads to a new way of thinking about a problem.” which is essentially equivalent to Koshland’s ‘Challenge’ discovery, whereas ‘small advances’ can be seen as a ‘Charge’. Hollingsworth’s definition does not address ‘Chance’ discoveries as they are serendipitous and therefore have no precursory research directly linked to the discovery.

This discovery-oriented approach of describing knowledge creation dynamics differs from philosophy of science ‘systems level’ approaches, such as Kuhn’s distinction (Kuhn, 1962) between ‘normal’ and ‘revolutionary’ science, the latter being characterized by ‘paradigm shifts’. Paradigm shifts start as a ‘localized’ diversion of an existing theoretical framework to explain observations that conflict with the theory. The consequence is that from a relatively small nucleus the new theory gains support of scholars resulting in growing numbers of publications. Paradigm shifts evolve gradually (Isnard and Zeeman, 1976) caused by the fact that scientists are reluctant to switch theories and concepts. During a paradigm shift increasing numbers of unique

⁹⁰See for instance van Andel (1994))

authors, institutions, journals, journal categories, and countries, where institutions are located, on publications related to the new paradigm can be observed.

Related studies, such as (Andersen et al., 2006) focus on the cognitive structure of scientific revolutions and the changes in concepts during paradigm shifts. (Worrall, 2003) discusses the various views on theory changes in scientific progress, whereas (Isnard and Zeeman, 1976) use catastrophe theory to explain the mechanisms behind the switching of views within societies.

5.3 Methodology

The various types of discoveries should manifest themselves by different types of events that can be externally detected through ‘bibliographical signals’. The solution for a long-standing scientific problem, for instance producing free standing graphene (Winnink, 2012), can lead to an impulse-like sudden increase in the number of research publications. To select relevant scholarly publications we used the topic ‘intron’ in the on-line version of Thomson-Reuters Web-of-Science database (WoS). This search resulted in 31,408 publications, articles and conference proceedings, for the period 1980–2012.

Patent publications were gathered by searching the EPO Worldwide Patent Statistical Database (PATSTAT), October 2011 version, for the terms “intron” and “non coding DNA” in titles and abstracts this resulted in 1,284 patent publications in 677 patent families for the period 1984–2012. We group patent publications describing the same invention in ‘simple patent families’ to prevent double counting⁹¹.

The data set contains intron-related publications from 1980 (Figure 5.1) onwards. The figure shows a very significant increase for scholarly publications in 1990–1991. To link scientific discoveries and technology we use the citing-cited links between patent-publications and scholarly publications. These links are called non-patent-literature references (NPL).

5.4 Results

General trends

Figure 5.1 shows, on a logarithmic scale, the trends for the numbers of scholarly publications and patent applications. The sudden increase in the number of scholarly publications from 1990 to 1991 stands out. We show that there are two breakthroughs of different type involved. One is the discovery of ‘introns’ in 1977, a ‘challenge’ breakthrough. The second is a ‘charge’ breakthrough marked by the sudden increase in intron-related scholarly publications from 1990 to 1991. The figure also depicts a continuous increase

⁹¹Because patents rights are national rights and patent procedures have several distinct phases patenting is a complicated process, in which several publications of the same invention can, and normally do, co-exist.

5.4 Results

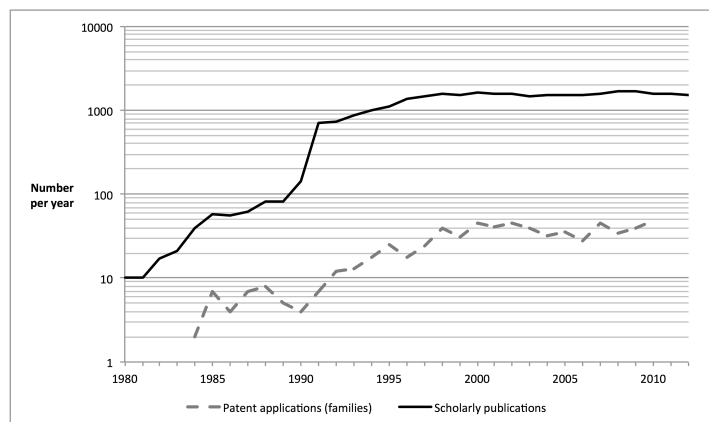


Figure 5.1: Trends in scholarly publications and patent applications

in patent applications from 1990 onwards until reaching a plateau level in 1998. The number of scholarly publications reaches an output plateau level at about the same time.

Figure 5.1 also reveals a low number of patent applications before 1992 indicating that intron related scientific knowledge was not yet used extensively in the development of new technologies. Figure 5.2 exhibits annual trends for the numbers of unique authors, journals, journal categories, organisations, and the number of countries where organisations that are ‘active’ in the intron-field at a certain moment are located; we presume that an actor that became active stays active. All these variables show a simultaneous and steady increase from 1980 onwards indicating that the theoretical concepts that form the basis for the selected scholarly publications spreads among the scientific community. These variables show the same ‘phase shift’ from 1990 to 1991 as in figure 5.1.

In an attempt to explain the sudden increase 1990–1991 we classified the scholarly publications⁹² into six different categories, each encompassing a different type of science according to the cognitive-institutional environment in which the research was done (Tijssen, 2010). We could classify 85% of the publications in the document set. Figure 5.3 shows the results. The sudden increase is especially marked in ‘application-oriented’ categories ‘Industrial/medical development’, ‘Clinically relevant science’ and ‘Industrially relevant science’. Our checks of the Web of Science indicate that this increase is not caused by new journals that were introduced into this bibliographical database. Although our inspection of the author affiliate addresses indicates that new institutional contributors (universities, research institutes, or other organisations) did enter the intron-field in 1991, the rise of publication output

⁹²This classification is developed at CWTS and linked with the in-house version of the Web-of-Science database (TR-CWTS WoS). Linking with the on-line version of the WoS is based on ‘Accession Numbers’.

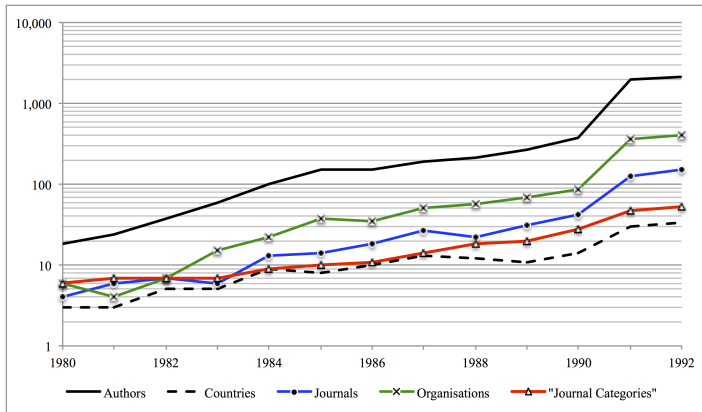


Figure 5.2: Frequency counts of unique authors, journals, journal categories, countries, and organisations that produced intron-related scholarly publications

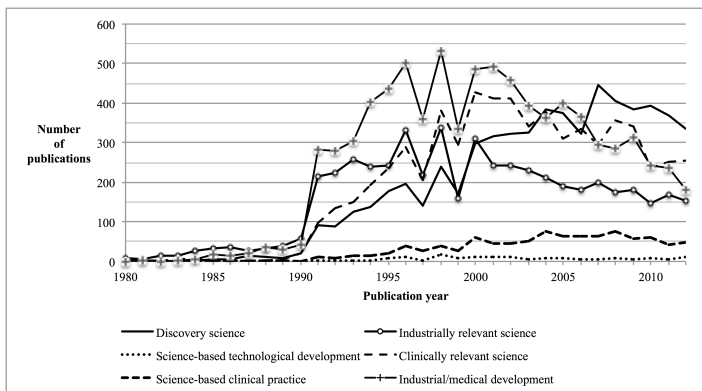


Figure 5.3: Trends in scholarly publications differentiated by type of science

occurs mainly within incumbent institutes, especially the US universities.

We constructed two subsets of scholarly publications; one for 1990 and one for 1991. Using the VOSviewer⁹³ we created two maps (Figures 5.4 and 5.5) showing the spatial configuration of keywords and phrases as they occurred in the titles and abstracts⁹⁴ of publications. Each of these content identifiers in these maps occurred in at least three publications. The same clustering and mapping algorithms were used to assure that the visible differences in the configuration reflect changes in content and relational structures.

⁹³For information on the VOSviewer see <http://vosviewer.com>

⁹⁴Before 1991 the Web-of-Science contained abstracts for only a fraction of the publications. We managed to obtain 138 abstracts for the 143 1990-publications, and 696 abstracts for the 703 1991-publications using additional sources.

5.4 Results

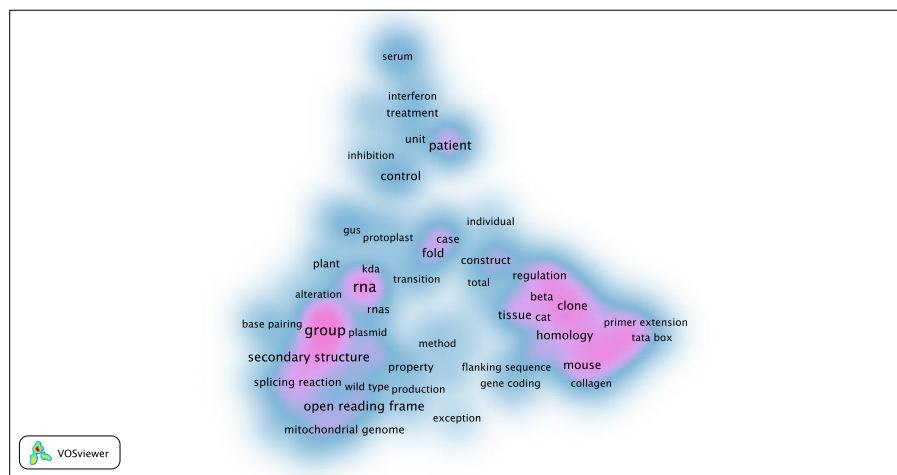


Figure 5.4: Term map for scholarly publications published in 1990

In general, maps using terms from relevant documents to illustrate the evolution of a research area in science look for successive years quite similar. The differences between the two maps reflect considerable changes from 1990 to 1991 (Figure 3). Close inspection of the keywords and phrases in the maps shows a shift towards application of scientific knowledge for technological development.

We find that 175 scholarly publications are cited within 112 patent families as non-patent literature (NPL) citations. These NPL citing-cited links are direct and visible connections between published science and patented technology. The network map in Figure 6 displays these NPL citing-cited links (citing-cited links between patents are omitted in order to highlight the science-technology interface) where the blue bubbles represent publications, while the white ones represent patent families. The graph shows a large number of dyads, mainly consisting of relations between a single publication and a patent. These are the multitude of ‘duplets’ that occupy the centre. This patent-publication network contains three major patent-publication clusters. To place these clusters on a time-line we use the earliest application date of a patent family as application year for the whole patent family. For publications cited three or more times (maximum is 6 citations) as NPL the identification numbers from the Web of Science (UT’s) are shown.

We used the patent classification codes assigned to the patent publications to determine the main ‘technological’ topic of a cluster. Based on the same scientific knowledge base (i.e. ‘introns’ science) the technologies represented in these three clusters diversify into three distinct technologies. ‘Recombinant DNA⁹⁵-techniques’ is the topic of Cluster 1 and the patent applications were applied for in the period 1992–2002. Cited research publications in this

⁹⁵DNA that has been formed artificially by combining constituents from different organisms

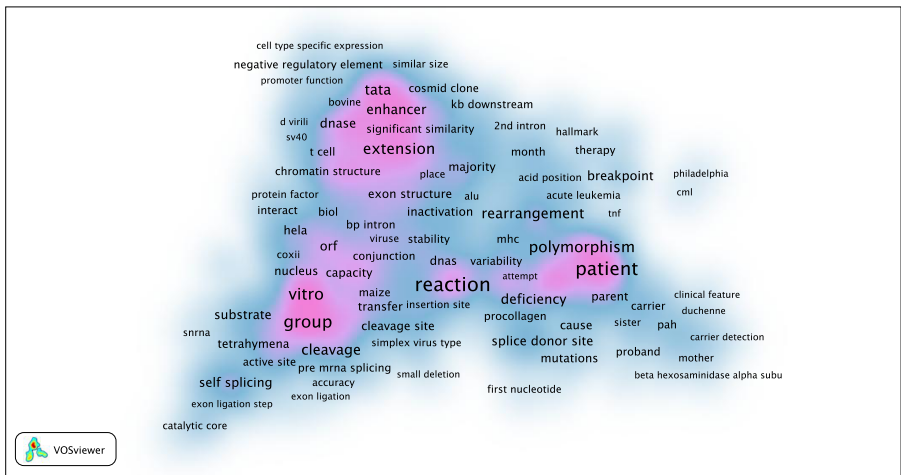


Figure 5.5: Term map for scholarly publications published in 1991

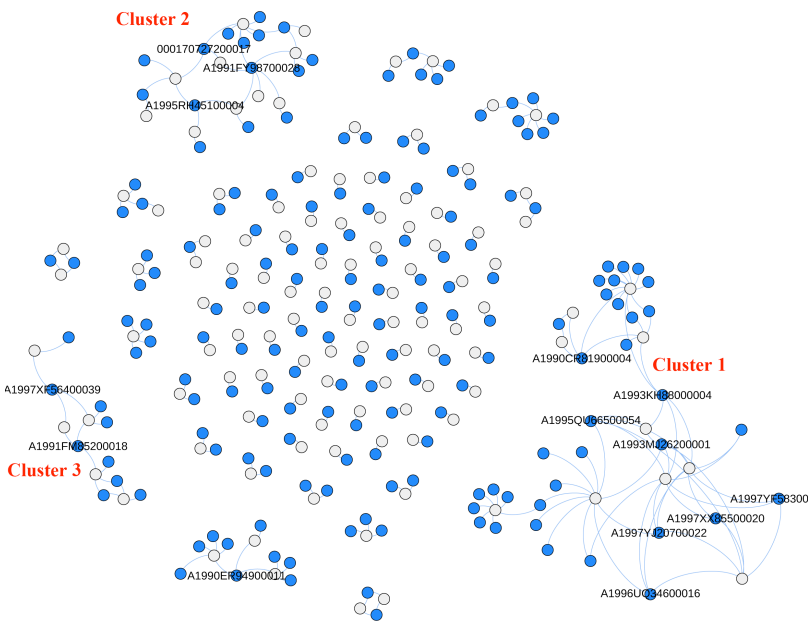


Figure 5.6: Patent-publication network according to non-patent-reference connections (1980-2010)

cluster were published between 1988 and 2005⁹⁶. Cluster 2 focuses on ‘DNA vectors’⁹⁷, patent applications in this cluster were applied for from 1998 until 2010, where the NPL-cited publications were published between 1991 and 2004. Cluster 3 covers the area of ‘Techniques to modify DNA to express or suppress genes’. The earliest patent application in this cluster is from 1995 and the most recent from 2005. The cited publications were published from 1986 until 2001.

5.5 Concluding remarks

A paradigm shift in science starts as a ‘localized phenomenon’, sometimes a discovery or breakthrough, the effects of which gradually spread throughout relevant research fields. This knowledge dissemination process reveals as simultaneous increases in numbers of unique authors, journals, journals categories, research organisations and the spread of different countries where those organisations are located. Figure 5.1 shows that the number of patent applications is low prior to 1992, indicating that scientific knowledge shortly after the breakthrough in 1977 is not sufficiently mature for applications in new technologies. It also illustrates that the discovery in 1977 was a ‘Challenge’ breakthrough.

The sudden increase in the quantity of scholarly publications in 1990–1991 marks the stage in which intron-related science becomes mature in the sense of relevance for new technological applications; the increase is mainly visible in application-orientated scientific areas and marked by application oriented terms. The citing-cited network between scholarly publications and patents reveals three major clusters of publications. Patent applications in these three clusters start to appear after 1991, thus signalling a direct link between science and technology. The three clusters present technological differentiation based on the same scientific knowledge.

The observations we made support the two main research questions we addressed in this study. We are able to identify and characterize a breakthrough, involving a paradigm shift, in terms of bibliometric variables and indicators, and identified, bibliometrically, the stage in the development process where scientific knowledge is being used for work on science-based technologies. We furthermore conclude that the link between scientific knowledge and patented technology is established after a ‘Charge’ discovery successive to a ‘Challenge’ discovery.

Further research will focus on transformative developments in intron-related research since the ‘challenge’ breakthrough in 1977 that led to the ‘charge’ breakthrough in 1990–1991.

⁹⁶Due to the complexity of the patenting process documents published after the filing of a patent application can show as cited non-patent-publication

⁹⁷An autonomously replicating DNA molecule into which foreign DNA fragments are inserted and then propagated in a host cell

Chapter 6

Validation study

The content of this chapter is equivalent to the manuscript:

Winnink, J.J. Tijssen, Robert J.W., and van Raan, A.F.J. (Submitted for publication). Can early-detection algorithms of breakout papers uncover scientific breakthroughs?.

The only differences between the article to be published and the text in this chapter are of typographic nature to assure that the layout is in line with the typographic design of this PhD thesis. The information presented in section 'Appendix 6.A Supplementary material' on page 125 will also be available on the website of the publisher in a separate file accompanying the journal article.

Abstract

In order to test our hypothesis "It is possible to design, develop, implement, and test an analytical framework and measurement model as a general-purpose tool using bibliographic information for early detection of breakthroughs in science?" we have conducted in the past few years a series of case studies to develop computerized algorithms. The algorithms facilitate the early detection of 'breakout' papers that emerge as highly cited and distinctive. These search & detection algorithms, combining computer-aided data mining with decision heuristics, assess structural changes within citation patterns with the international scientific literature. The citation impact time window in this study is 24–36 months after publication of research papers. The wider applicability of these algorithms, across all science fields, had not yet been ascertained. In this paper we report on our test results, in which five algorithms were applied to the entire Web of Science database, i.e., to all research articles from the period 1990–1994 presenting original scientific or technical research. We succeeded in detecting many 'breakout'

papers with distinctive citation impact profiles. A small subset of these breakouts is classified as 'breakthroughs': Nobel Prize research papers. Furthermore papers occurring in Nature's Top-100 Most Cited Papers Ever, or papers that are still (highly) cited by review papers, by patents, or frequently mentioned in today's social media were found. We also compare the outcomes of our algorithms with the results of the breakthrough-detection algorithm developed by Redner in 2005. The detection rates of the algorithms developed in this study vary but overall they present a superior tool for tracing breakout papers in science, when compared to methods that select the world's most highly cited 'hot papers'. For the final conclusion, if an identified early-stage breakout paper presents a 'breakthrough' in science, the opinion of subject experts is needed.

6.1 Introduction

6.1.1 Theoretical and conceptual issues

In this study we try to prove the validity of our hypothesis "It is possible to design, develop, implement, and test an analytical framework and measurement model as a general-purpose tool using bibliographic information for early detection of breakthroughs in science?". Scientific discoveries that have an above average impact on science are often called 'breakthroughs'. As there is no strict and general accepted definition of the term 'breakthrough' characterizing a scientific discovery, as a breakthrough is not a straightforward process. In order to understand the evolution of science various theoretical models have been proposed. Bettencourt and colleagues (Bettencourt and Kaiser, 2011, 2015; Bettencourt et al., 2009) propose a percolation model. Bonaccorsi and his co-authors focus on dynamic patterns in the evolution of science (Bonaccorsi, 2008), and on the way new fields of science come into existence (Bonaccorsi, 2010; Bonaccorsi and Vargas, 2010). In (Chen et al., 2009) an explanatory and computational theory of transformative discoveries in science is proposed. The use of disease-propagation models to describe knowledge diffusion is discussed in Cintron-Arias et al. (2005). Perla and Carifio (Perla and Carifio, 2005) use a catastrophe theory representation to depict and develop a formal nonlinear model of scientific change in concordance with Kuhn's hypotheses (Kuhn, 1962). Experiments play a crucial role in formulating an explanation of the RNAi anomaly as is explained in (Sung, 2008). Population dynamic models can be used to explain the diffusion of ideas among scientists (Vitanov and Ausloos, 2012).

These and other theoretical models describe the behaviour of science as a system; linking the evolution of the science system with individual publications remains a challenge. The purpose of this study is to identify at early stage areas that evolve into 'hot spots' in science. Science is not a gradual evolving system but a system that alternates between incremental advancements and major high-impact discoveries. Kuhn (Kuhn, 1962) calls these two states 'normal science' and 'revolutionary science'. In the latter case a

'paradigm shift' is involved, and in the first case not. High-impact discoveries represent major steps forward. Identifying such sudden high-impact changes is complicated as it takes time before the conclusion can be drawn a discovery has a major influence on future science, and secondly because there is not a well defined general-accepted objective definition for a breakthrough. In this study the approach is taken to analyse citation profiles of scholarly papers that present scientific discoveries classified by subject experts as a 'breakthrough'. These citation profiles are searched for characteristic breakthrough-signalling patterns. Such breakthrough-signalling patterns are then used as the basis for computerised automatic algorithms. In this way the need for a strict definition of the term breakthrough is circumvented and replaced by a more pragmatic definition based on expert opinions. The focus is herewith on the impact of an individual paper on the evolution of science. Our research question is answered by searching from multiple perspectives for scholarly publications that mark a major system change in the science system. We think the approach using a small number of cases studies to base our algorithms on is sound as Bettencourt and colleagues theorize that there is a universal character in discoveries and inventions (Bettencourt et al., 2009, p.220), and argue that circumstantial evidence found in several other studies, such as (Gerstein and Douglas, 2007; Leskovec et al., 2005; Uzzi and Spiro, 2005), supports this theory. If such a universal character does exist it opens up the possibility to analyse individual discoveries, as we do, that are recognised by experts as a breakthrough and search the bibliographic data for characteristic patterns. Such patterns should result in generally applicable computerised search algorithms to identify – at an early stage – those scientific papers in the open scholarly literature that signify scientific discoveries. We call scholarly publications identified by our search algorithms 'breakout' papers.

Our research focuses on constructing algorithms to improve the understanding of discoveries in science, and more particularly in scientific breakthroughs. Early stage identification of breakout papers could help in finding general mechanisms related to scientific discoveries. Bettencourt et al. (2009, p.219) assert "... It has long been a goal of the history, philosophy, and sociology of science and, more recently, of bibliometrics and a new 'science of science', to identify (quantitative) indicators or circumstances that reveal moments of scientific and technological discovery ...". These authors further state "... Thus, a more quantitative 'science of science' may allow society to reap the benefits of new discoveries sooner ...". We advance the following hypothesis:

"It is possible to design, develop, implement, and test an analytical framework and measurement model as a general-purpose tool using bibliographic information for early detection of breakthroughs in science."

To find out if this hypothesis can be validated we have conducted in the past few years a series of case studies to develop computerized algorithms. By

running a series of algorithm-driven large-scale systematic searches on citation patterns, we aim to identify research papers that mark a major change in the science system; we call these ‘breakout papers’. These papers are likely to have an above-average impact on follow-up scientific literature, as marked by a relatively large quantity of citations that accumulated rapidly after the publications appeared in the public domain. Research papers that achieve rapid impact are usually cited within months after publication, and certainly within two or three years.

Discoveries and inventions often come in a manifold, according to Ogburn and Thomas (1922); they may have different appearances, even if only slightly different; they emerge at various moments, or appear at different places. De Solla Price (Price, 1963, p.65–66) discusses this ‘multiplicity of discoveries’ while citing Merton Merton (1961) who confirmed Ogburn’s observations, and linking multiplicity to Kuhn’s (Kuhn, 1962) concept of ‘normal science’ in which discoveries are ‘expected’ outcomes of ‘evolutionary’ scientific progress. Scientific discoveries come in several flavours based on the impact a discovery has on this evolution. Only a small number of scientific discoveries lead to major structural changes in science – at the level of research fields or disciplines – and providing novel insights and opening avenues for further productive research. Such rare discoveries are often referred to as ‘breakthroughs’, or by synonyms such as ‘step forward’, ‘quantum leap’, ‘revolution’, and other variations. Subject experts are able, with the benefit of hindsight, to assess the impact of a particular breakthrough on evolutionary processes in science. Major general and interdisciplinary journals like National Geographic, Nature and Science, regularly publish overviews of what they perceive as the major scientific discoveries in a preceding period. These lists are usually based on expert opinions and do not specify what exactly is meant by a breakthrough or major discovery. And for a good reason: there is, as mentioned before, no generally accepted description, let alone a universal one, of the term breakthrough (or synonyms) that can count on full support throughout the scientific community.

6.1.2 Classification and identification of scientific discoveries

In order to pragmatically operationalize and classify scientific discoveries, in a way that enables large-scale systematic searches across worldwide research literature, we rely on the conceptual work by Hollingsworth (Hollingsworth, 2008, p.317) who argues “... A major breakthrough or discovery is a finding or process, often preceded by numerous small advances, which leads to a new way of thinking about a problem ...”. Secondly, we follow Koshland’s ‘Cha-Cha-Cha’ theory (Koshland, 2007, p.761) in which three main types of discoveries are distinguished:

“... In looking back on centuries of scientific discoveries, however, a pattern emerges which suggests that they fall into three categories —Charge, Challenge, and Chance— that combine into a ‘Cha-Cha-Cha’ Theory of Scientific Discovery ...”

According to Koshland: “Charge discoveries solve problems that are quite obvious ...”, while “Challenge discoveries are a response to an accumulation of facts or concepts that are unexplained by or incongruous with scientific theories of the time ...”, where “Chance discoveries are those that are often called serendipitous ...”. As for the latter category, Koshland states: “... not only would include Pasteur’s discovery of optical activity (D and L isomers), but also W. C. Röntgen’s X-rays and Roy Plunkett’s Teflon® ...”. When framed with Thomas Kuhn’s general theory of scientific progress, these Charge discoveries are more likely to occur during periods of ‘normal’ science, whereas Challenge and Chance contribute to ‘revolutionary’ science, which is marked by paradigm-shifting breakthroughs and major epistemological changes within fields of science.

The occurrence of a breakthrough is hard to foresee (Bettencourt et al., 2009). Even within weeks or months, and sometimes many years, after the fact, the significance of a major scientific discovery is not always clear or obvious. Ball Ball (2004) states that in general systems, such as in this case research fields and fields in science, need a certain ‘critical mass’ to undergo a major change. As mentioned before Bettencourt and co-authors (Bettencourt and Kaiser, 2015) conclude that it is possible to develop a general theory and detailed models that govern the formation of fields in science. Recent publications (Scheffer, 2010; Scheffer et al., 2009) show that information transmitted by ‘dynamic systems’ may contain specific patterns that precede and signal an upcoming phase-shift or transition to a new stable state, possibly at meta level. Rather than trying to find such early-warning signals, our research objective had a more modest goal: we focus our attention on detecting and analysing citation patterns of papers in the world’s scientific literature across a two or three year time-span. However, even in retrospect, the precise moment of a breakthrough is often impossible to pinpoint on any time scale. We therefore take the publication date of the seminal research paper as time reference in order to determine a discovery’s entry into the science system.

6.1.3 Research framework and detection algorithms

Various types of literature search-methods exist to track and trace high-impact papers. High-impact papers in this context receive an above-average number of citations when compared to other publications in the same field and time period. Without additional information, particularly the opinion of subject experts, these methods are unable to conclude whether a high-impact paper really presents a breakthrough in science as the information provided by these methods is in principle first evidence based on quantitative considerations. This unavoidable restriction holds for methods like the algorithm presented in (Redner, 2005), in which a set of citation counts is introduced that, in Redner’s opinion, identify ‘breakthroughs’ in science, and for the methods used by Schneider and Costas (2014). The algorithms come up with what we dub ‘breakout’ papers. Thomson Reuters’ (TR) concept of ‘hot papers’ which represents another early detection method. These ‘hot’ pa-

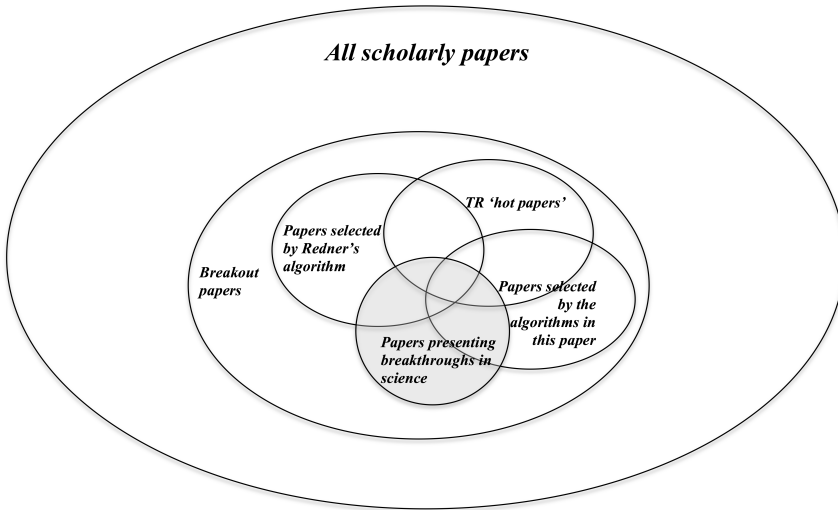


Figure 6.1: Overview of conceptual and analytical framework

pers are selected by virtue of being among the top 0.1% of the most highly cited papers in a current bimonthly period within their field of science (see <http://archive.sciencewatch.com/about/met/core-hp/>). Most of these papers eventually qualify as ‘citation classics’, i.e. they continue to be cited significantly over time. TR identifies hot papers in four of the areas (Physiology or Medicine, Physics, Chemistry, and Economics) for which Nobel Prizes are awarded. This information is supplemented with information about the ‘Nobel Prize history’ for both the papers and the authors (Pendlebury, 2015). TR publishes on its Science Watch website (<http://ScienceWatch.com>) - an open Web resource for science metrics and research performance analysis - under the heading ‘Predictions’ what TR sees as candidates that might be awarded a Nobel Prize.

Ponomarev et al. (2014b) studied the dynamics of fast-growth citation impact. A key question they asked is whether the use of the interdisciplinarity in scientific fields of the referenced and cited papers, or the geographical diversity in the affiliations of the authors of citing papers, could improve the identification of ‘breakthrough paper candidates’ (Ponomarev et al., 2014a). All these detection algorithms are designed to identify breakout papers. Some methods focus on identifying papers as quickly and as reliably as possible, after their publication, as breakout papers. The underlying implicit assumption is that only a small subset of these papers are likely to be generally accepted scientific ‘breakthroughs’. Our ‘early warning algorithms’ also fit in this citations-based analytical approach. Clearly, the various algorithms may identify the same breakout papers. We will test this assumption by comparing the results of our set of algorithms with those of Redner’s algorithm (Redner,

2005). Figure 6.1 presents a stylized overview of the conceptual framework, and shows possible overlap between the search results produced by the various detection methods.

In this study we test five algorithms to identify breakout papers at an early stage. These algorithms were developed as outcomes of a series of four case studies in which we analysed the citation patterns of scientific discoveries that subject experts regarded as breakthroughs. These studies were on: HIV/AIDS medicine (Winnink and Tijssen, 2014), Graphene (Winnink and Tijssen, 2015), Introns (Winnink et al., 2013), and Ubiquitin-mediated proteolytic system (Winnink et al., 2015). In each case we went beyond straightforward ‘hot papers’ citation counting; we searched systematically for any kind of remarkable change in those citation patterns that characterised the underlying breakout papers.

How generic are these algorithms in terms of their efficacy across all fields of science? This paper presents the test results, focusing on three research questions.

- Can the algorithms be used as a generally applicable method to identify breakout papers, and if so under what data availability conditions?
- What are the similarities and differences between the algorithms in terms of their ability to detect breakouts?
- Can we determine the effectiveness of each algorithm in terms of identifying breakout papers that are generally regarded as breakouts and potential breakthroughs?

The next sections describe and interpret the main findings of our study, and the general conclusions one can draw. Detailed information concerning the results of this research is added to this manuscript as Supporting Information. References from the main text to tables and figures in the supporting information are Table 6.10, Figure 6.3 etc.

6.1.4 Data source and methods

Our bibliographic database consists of research papers extracted from CWTS’s in-house offline version of Thomson Reuters’ Web of Science database (WoS). We selected all 2,715,651 papers published in the period 1990–1994 that were tagged with the WoS document types ‘article’ or ‘letter’. These documents are most likely to report on ‘original research’. We opted for the time period 1990–1994 to track the effect of a discovery over an extended period of time, and to verify and validate whether selected papers are currently – in retrospect – (still) regarded as breakouts or breakthroughs. For reasons of citation impact normalisation, we adopt two publication-based delineations of scientific disciplines: (1) ‘Categories’, the equivalent of Thomson Reuters’ subject categories – information on subject categories can be found at ‘<http://ip-science.thomsonreuters.com/cgi-bin/jrnlst/jlsubcatg.cgi?PC=D>’, and (2) ‘Clusters’ derived from a citation-based clustering algorithm developed at CWTS (Waltman and van Eck, 2012); we refer to this method as the ‘CWTS document clustering method’. Each of the 251 Categories comprises a set of

entire WoS-indexed journals; the 865 Clusters each consist of large numbers of individual research papers. In line with other scholars we assume that each these disciplines is: "... an area of science consisting of the following elements: a central problem, a domain consisting of items taken to be facts related to that problem, general explanatory factors and goals providing expectations as to how the problem is to be solved, techniques and methods ..." (Darden and Maull, 1977).

To narrow down our search, we selected those papers that belong to the top 10% most highly cited during the first 24 months after publication per Category or Cluster per year. The Categories subset contains 253,558 highly cited papers, Clusters subset is comprised of 214,827. Table 6.1 presents summary statistics of both datasets. All computations and analyses were carried out separately on both datasets.

6.1.5 Breakout detection algorithms

Our algorithms meet the following general specifications: (1) directly derived from data-analytical results in our case studies; (2) systematically applicable across the Web of Science; (3) systematically applicable at the level of individual research papers; and (4) their implementation does not require any special pre-processing of WoS-based bibliographical data. We developed five 'general purpose' algorithms, each representing a specific characteristic of citation impact patterns:

- Application-oriented Research Impact (ARI): citations to the breakout paper come in majority from papers within the application-oriented domains of science (application-science papers), whereas the breakout paper is in the discovery-science domain (discovery-science papers);
- Cross-Disciplinary Impact (CDI): citations to the breakout papers come from an expanding number of science disciplines;
- Researchers-Inflow Impact (RII): influx of new researchers citing the breakout paper;
- Discoverers-Intra-group Impact (DII): citations to the breakout paper are concentrated within papers mainly produced by members of the same research group that produced the breakout paper;
- Research-Niche Impact (RNI): the breakout paper and its citing papers are tightly interconnected within a small research domain.

Further details of each algorithm are presented below. ARI, CDI and RII are more likely to identify Charge discoveries (i.e. solving well-known and well-defined problems – Kuhn's *normal science*), while DII and RNI are better equipped to find Challenge discoveries (i.e. explaining strange, unexpected phenomena – Kuhn's *revolutionary science*). As for Chance discoveries and breakthroughs, given their random nature, we discarded the search for generally applicable algorithms that may systematically identify such cases within a short time-span.

ARI algorithm (Application-oriented Research Impact)

The purpose of this algorithm is to identify papers that describe new knowledge on the boundary of ‘discovery-oriented science’ and ‘application-oriented science’ (Tijssen, 2010). The algorithm emerged from the case study in which we noticed remarkable shifts over time in the ratio of citations from discovery-science papers and applied-science papers in the field of Introns (Winnink and Tijssen, 2013). The focus is on papers having a substantial list of references and are highly cited within the first 24 months after publication. The majority of the referenced papers focus on ‘discovery-oriented science’, whereas the citing publications focus mainly on ‘application-oriented science’. Each breakout paper should meet the following selection criteria that are based on all papers in the Category and Cluster document sets:

- Number of cited papers ≥ 30 , this is the lower boundary for the top decile of the number of original-research papers in the reference lists;
- Number of citing papers within 24 months ≥ 49 , this is the lower boundary for the top decile of the number of citations received within the first 24 months by the most-highly-cited papers;
- Number of citing papers $>$ number of cited papers;
- Majority of the cited papers focus on ‘discovery-oriented science’;
- Majority of the citing papers focus on ‘application-oriented research’.

CDI algorithm (Cross-Disciplinary Impact)

This algorithm captures the diffusion of citing sources among multiple research disciplines. We expect to find breakout papers that are cited by an increasingly larger number of disciplines over time. The level of cross-disciplinary impact is defined as the number of different disciplines (either Categories or Clusters) that are assigned to each of the citing papers. Given the more homogenous disciplinary composition of each Cluster, as compared to each Category, one would expect less interdisciplinary citation flows between Clusters. This aspect is especially noticeable during the first few years after publication, as is illustrated in the support information in Figure 6.6 and Figure 6.7. Breakout papers meet the following lower threshold values per citation time window, and that are based on the values for the ‘Hazuda paper’ (Hazuda et al., 2000), which is central in our case study of HIV/AIDS research (Winnink and Tijssen, 2014):

- Categories: 1 year: >9 citing disciplines; 2 years >17 disciplines; 3 years >24 disciplines;
- Clusters: 1 year: >2 citing disciplines; 2 years >5 disciplines; 3 years >8 disciplines.

RII algorithm (Researchers-Inflow Impact)

Our case study of Graphene research (Winnink and Tijssen, 2015) identified papers that attract a remarkable increase in different citing researchers. Here we expect to identify breakout papers that have an impact on an increasingly

large community of research-active scholars in the research domain. Focusing on the annual number of these unique authors, who are first-authors on citing research papers, we measure the inflow rate by comparing the increase in the number of researchers at the end of the 1st year after publication, and at the end of the 3rd year. Selected papers should show an increase of at least 52 new citing first-authors. This threshold results from the increase in new citing first-authors between the end of the 1st year after publication, and at the end of the 3rd year for the paper on the Graphene discovery (Novoselov et al., 2004), which was central in the analysis shown in (Winnink and Tijssen, 2015).

DII algorithm (Discoverers-Intra-group Impact)

In our study of Ubiquitin research (Winnink et al., 2015), we find that the breakout papers that describe the breakthrough received most of its citations, within the first two years, from papers (co) authored by authors from the same ‘core group’. The discovery is at first predominately recognized and built upon by members of the same group. This algorithm is designed to find breakout papers where many citations are from papers with authors that share co-authorship relationships with the cited authors. The following selection were applied:

- 90% of the citations are ‘within-group’ citations;
- Within-group papers are defined as papers of which at least 66% of the authors belong to the core group. This specific lower threshold avoids the inclusion of those papers for which only one member of a small group — 3 or 4 members — is (co) author;
- The minimum size of a core group is three, which value is chosen to guarantee that in combination with the above-mentioned 66% threshold, only papers written by at least two authors of the core group are considered;
- Citations are tracked within the first two years after publication.

RNI algorithm (Research-Niche Impact)

This algorithm, also originating from the Ubiquitin case study, searches for sets of citing and cited papers, within Categories or Clusters, with above-average rates of citation-interconnectedness. A breakout paper creates a ‘citation knot’, i.e. a set of papers that cite the breakout paper but also cite at least one ‘auxiliary’ paper with direct citation ties to the breakout paper. This closely-knit set of citing and cited papers represents a ‘research niche’. The next threshold values are determined by analysing for the period 1980–1982 the network of papers citing the two breakthrough papers from 1980 that in conjunction describe the ubiquitin discovery (Winnink et al., 2015).

- The number of citations received by the breakout paper, within this niche and within the first year, is larger or equal to three times the number of interconnected papers within a citation cluster;

- The lower threshold for the number of breakout-related papers in the ‘citation knot’ is 8.

6.2 Robustness of the algorithms

We define ‘robustness’ as the ability of each algorithm to identify the same breakout paper(s) irrespective of the total number of citations a paper received within two years. We tested the robustness empirically by implementing citation count thresholds of 1, 2, 4, ..., 1024 citations. In Table 6.1 we present a subset of the findings for threshold values of 1, 4, 16, 64, and 256 citations. For all threshold values the results are presented in Table 6.13 and Table 6.14.

The RII and CDI detection algorithms manage to capture many breakouts, and are most effective for both datasets (Categories and Clusters) because these algorithms focus on the more frequently occurring of discovery type ‘Charge’. The CDI rates are much higher in Clusters because the CWTS document-clustering method groups documents together on the basis of citation relations. These document-clusters may contain papers from multiple WoS subject categories; this means that diversity is in fact already achieved within a cluster, thereby reducing inter-cluster relations. The consequence is that a different and lower threshold level is used when CDI is applied to Clusters. In the long run, this ‘vanishing diversity’ effect largely disappears, as is shown in Figure 6.6 and Figure 6.7.

ARI and especially RNI are much more targeted towards rarer types of breakouts, because ARI focuses on breakouts that bridge the gap between discovery-oriented science and more application-oriented science. The focus of RNI is on areas where the fabric of the citation network is more dense. DII sits between these extremes, but is by far the most threshold-sensitive algorithm, within both Categories and Clusters; it ceases to be effective above the threshold of 16 citations. By virtue of their search criteria, DII and RNI work best within social networks and micro research areas with low-citation levels.

In contrast, the breakout hit rate of RII is only affected by higher (≥ 64) values of the threshold, which follows directly from the requirement that in order to be selected as a breakout paper, it has to be cited by at least 52 papers within 24 months; this high threshold for RII is explained above. The performance of RNI is only slightly threshold sensitive until a threshold of 64 citations. RNI is a very selective algorithm as it searches for sets of citing papers with relatively large numbers of cross-citation relationships. The results are identical for Categories and Cluster, and decrease above the threshold of 64 citations within 2 years. ARI selects four times more papers in Categories than in Clusters. In the section on ‘Discussion and overall findings’ we give a possible explanation for this phenomenon.

CDI-generated hit rates are significantly affected within Clusters, although the number of identified breakouts remains large, because of the already dis-

Table 6.1: Robustness of algorithms

Categories	Lower threshold for the number of citations within 24 months			
	1	4	16	64
Total number of papers above threshold	253,558	238,009	53,014	2,996
Number of breakout papers detected				
Application-oriented Research Impact (ARI)	264	264	204	0
Cross-Disciplinary Impact (CDI)	1,276	1,276	1,275	924
Researchers-Inflow Impact (RII)	3,543	3,543	3,543	2,891
Discoverers Intra-group Impact (DII)	577	410	2	0
Research-Niche Impact (RNI)	19	19	19	13

Clusters	Lower threshold for the number of citations within 24 months			
	1	4	16	64
Total number of papers above threshold	214,827	201,514	51,666	2,926
Number of breakout papers detected				
Application-oriented Research Impact (ARI)	60	60	60	0
Cross-Disciplinary Impact (CDI)	13,477	13,477	12,096	2,320
Researchers-Inflow Impact (RII)	3,501	3,501	3,501	2,857
Discoverers Intra-group Impact (DII)	674	483	2	0
Research-Niche Impact (RNI)	8	8	8	7

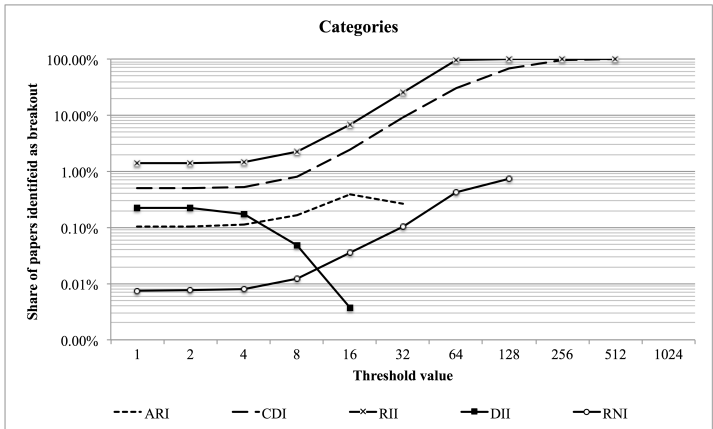


Figure 6.2: Share of papers recognised as breakout as a function of the threshold value applied on Categories

6.2 Robustness of the algorithms

Table 6.2: Performance of the algorithms on the two datasets

	Number of breakout papers identified	of which matched by one algorithm	of which also matched by one or more of the other algorithms
Categories			
<i>Total</i>	4,946		
ARI	264	99.6%	0.4%
CDI	1,276	21.2%	78.8%
RII	3,544	71.4%	28.6%
DII	577	99.8%	0.2%
RNI	19	31.6%	68.4%
Clusters			
<i>Total</i>	15,074		
ARI	60	50.0%	50.0%
CDI	13,477	78.9%	21.1%
RII	3,544	20.8%	79.2%
DII	674	100.0%	0.0%
RNI	8	12.5%	87.5%

cussed way the datasets are constructed. In all, RII is robust up to a threshold value of 32 citations, and for CDI and RNI the robustness starts to break down at a threshold value of 16 citations. Beyond this threshold value the hit rates start to decrease (see Table 6.13 and Table 6.14). For ARI this hit-rate breakdown starts for Categories at a threshold value of 8, but for Clusters at 16 – the same value as for CDI and RNI. The DII algorithm should be considered not to be robust, as its hit rates already start to decrease at a threshold value of 2 citations.

As an indication of the performance of the algorithms, we calculated on the basis of both datasets for each algorithm the number of papers recognized uniquely by an algorithm as well as the number of papers recognized by multiple algorithms. Table 6.2 presents the results of these calculations. We observe that, except for RII, the performance of the algorithms varies for the datasets when measured in absolute numbers of breakout papers. This table also shows the ability of each of the algorithms, regardless of the dataset, to select papers that are not selected by any of the other algorithms that we developed. Because papers can be selected by multiple algorithms the total count is not an add-up of the counts for the individual algorithms.

Figure 6.2 shows the performance of the algorithms as a function of the threshold values applied on Categories. For Clusters this behaviour is visualised in Figure 6.11. Both CDI and RII select above a threshold value of 128 almost all documents, and reach the 100%-level for higher threshold values.

6.3 But is it a breakthrough?

As explained above, there is no objective measure to qualify or classify a scientific discovery, or its underpinning papers, as a breakthrough. Concepts or criteria from information science cannot be used because there is no straightforward or transparent heuristics for decision-making. One has to rely on assessments based on expert opinion and therefore accept a degree of subjectivity. Various assessment methods, each with relatively high levels of inter-rater reliability, offer guidance. The following additional verification metrics were used:

1. Papers documented to be relevant for awarded Nobel Prizes. If a Nobel Prize in physics, chemistry, or physiology or medicine is awarded for a single discovery or invention it considered a 'breakthrough'. The single publication or group of closely related publications in which such a discovery is presented signal this breakthrough;
2. Occurrence of a paper on Nature's 'Top-100 list of papers most cited ever' (van Noorden et al., 2014). The papers appearing in this list of publications are considered by the scientific community of particular importance. Not all papers on this list are scientific breakthroughs as is mentioned in one of the comments to this Top-100 list (Padhi, 2014). Although these papers do not display breakthroughs in science by definition, experts are able to judge if a paper presents a scientific breakthrough;
3. The number of times a paper is cited in WoS-indexed review papers; review papers provide an overview of the developments that occurred in a topical field of science over a certain period of time. Of all scholarly papers of types articles and letters published between 1990 and 1994 review papers cite 50%; this equals 62% of the papers that are cited at least one time. Special attention should be given to papers that are highly cited for instance papers belonging to the Top 1% (Table 6.4);
4. The number of times a paper is cited in worldwide social media (2012–2014). We conclude that a breakout paper stands out when it is still cited in social media 20 years after publication (1990–1994). Such scholarly papers should be at least looked at to see if they are really special;
5. The number of times a paper is cited in patents. Scholarly papers cited in patents bare relevance to the invention described in the patent, and are part of the scientific basis for the developments in a field of technology. In total 6% of the scholarly papers are cited in patents. Based on the number of times cited by patents 11 out of the 60 papers from the 'test set' belong to the top 2% percentile. Citations to scholarly publications in patents link the two domains 'science' and 'technology'. From these citations alone the conclusion can be drawn that the paper is interesting enough to be classified as a *break out* publication.

Awarded Nobel Prizes

Perhaps the best known of these are the annually awarded Nobel Prizes for internationally outstanding scientific achievements. We found eight awarded Nobel Prizes where scholarly work published between 1990 and 1994 was seen by the Nobel Prize committee as being of seminal importance. Five of those cases involve at least one of our identified breakout papers, now verified as a 'breakthrough' paper. Detailed information is presented in Table 6.16.

Nature's 'Top-100 list of papers most cited ever'

As a general frame of reference to assess the performance of the algorithms, we use Nature's 'Top-100 list of papers most cited ever', published in 2014 by Van Noorden and co-workers (van Noorden et al., 2014), which provides a list of the 100 most cited papers of all times. The top 100 include 13 of our breakouts (Table 6.15). The bibliographic information for these papers is presented in Table 6.18. Two of those 13 papers (Laskowski et al., 1993; Moncada et al., 1991) are not included in our tests because of their document types 'software review paper' respectively 'review paper', which were excluded from our analysis.

Additional verification metrics

We apply three additional methods, numbers 3–5 in the list discussed in the foregoing section, to help verify our identified breakout papers - all are based again on citation impact. But now these citations are from sources other than the WoS: patents and social media. For reasons of resource constraints our verification study cannot be applied to the full set of breakouts, but was done within a small sample of breakout papers. These were selected by applying the five algorithms to the two datasets (Clusters and Categories) separately. From each of these 10 applications, we selected the top-10 most cited papers in terms of citation count frequencies. This test set included 60 unique papers (40 of the 100 papers occurred more than once), of which 25 occur both in Categories and in Clusters, 20 exclusively in Categories, and 15 are found only in Clusters. The bibliographic information of these papers is provided in Table 6.19.

Table 6.3 presents the test results, highlighting the ability of the RII and CDI algorithms to identify Nature's Top 100 most-cited publications. More importantly, all these breakouts were also cited in at least one review paper. Patents also cite more than half of all breakouts detected by the RII, CDI and RNI algorithms, thus giving an indicator of the technological impact of the scientific discovery. These three algorithms also captured breakouts that generate, or still generate, a wider societal impact, when measured on the basis of social media ('altmetrics') for the years 2012–2014. The CWTS' social media database contains social-media data related to Internet blogs, news, Twitter and Facebook messages collected from the altmetric database

Table 6.3: Performance of algorithms within the test set

	Breakout papers	of which in Nature's Top-100 list	of which cited in review papers	of which cited in patents	of which cited in social media
ARI	15	0	15	7	0
CDI	33	11	33	25	10
RII	26	8	26	20	9
DII	16	0	16	2	0
RNI	12	0	12	6	4

provider Altmetric.com (<http://www.altmetric.com>) The two 'large-output' algorithms in Table 6.3 (RII and CDI) manage to produce the largest number of verified breakouts.

The algorithms developed in this study are constructed on the basis of the outcomes of case studies. One of the criteria used to select cases was that the scientific breakthrough discoveries resulted in new technological development, as shown by the occurrence of citations from patents. In this way the algorithms may contain in an implicit form a link between science and technology that could explain the occurrence of patent citations. Table 6.4 displays characteristics of the breakout papers identified by each algorithm. The detailed results are presented in Table 6.17. In Table 6.18 we comment on all 60 papers in the test set, and present the bibliographic information for these papers in Table 6.19.

Applying each of the algorithms to the test set results in two groups of documents for each algorithm. One group contains the papers that are selected (breakout papers), and the other group the papers not selected (non-breakout papers). To search for differences in the characteristics of the documents in both groups, the share of papers belonging to the top 3% percentiles is used. As the often-used top 10% percentile did not show differences in behaviour between breakout and non-breakout papers, we chose to use the top 3% percentile. These top 3% percentiles are based on the distribution of the number of citations received from the different sources by all papers (letters and articles) published in 1990–1994 that are covered in the WoS database; Table 6.5 shows the results.

Redner's algorithm

We implemented the algorithm to identify breakthrough papers published by Redner (2005), and use it as another benchmark for our algorithms. Redner argues the choice of the parameters as "...I arbitrarily define a discovery paper as having more than 500 citations and a ratio of average citation age to publication age of less than 0.4 ...". The concepts 'average citation age' and 'publications' age are best illustrated with the following example. A publication published in 2000 is cited 1 time in 2005 and two times in 2010. The average citation-age is in this case 8.33 years: $\frac{(1 \times 5) + (2 \times 10)}{3}$, because 1 cita-

6.3 But is it a breakthrough?

Table 6.4: The citation impact of the breakout papers based on the number of citations received from three different citing sources. The classification is based on the distribution of the numbers of citations received by all papers (letters and articles) published in the period 1990-1994 and covered in the WoS database.

Number of breakout papers		Citation impact (Percentile boundary)				
		Not cited in source	Low ($\leq 70\%$)	Medium (70-97%)	High (97-98%)	Very high ($\geq 99\%$)
Citing source: patents						
ARI	15	8	3	3	1	0
CDI	32	8	2	9	5	8
RII	26	6	2	7	4	7
DII	16	14	1	1	0	0
RNI	12	6	1	3	1	1
Citing source: review papers						
ARI	15	0	0	1	5	9
CDI	32	0	0	0	1	31
RII	26	0	0	0	1	25
DII	16	1	5	9	1	0
RNI	12	0	1	0	1	10
Citing source: original research papers						
ARI	15	0	0	0	2	13
CDI	32	0	0	0	1	31
RII	26	0	0	0	0	26
DII	16	0	2	13	1	0
RNI	12	0	0	0	0	12

Table 6.5: Characteristics of the breakout and non-breakout papers in the test set as a percentage of papers that belong to the top 3% percentile for all papers (articles and letters) based on the number of citations received from the different sources.

	Cited by review papers		Cited by patents		Times cited within 24 months
	Breakout papers	Non-breakout papers	Breakout papers	Non-breakout papers	Breakout papers
ARI	93%	64%	7%	27%	100%
CDI	100%	39%	41%	0%	100%
RII	100%	50%	42%	6%	100%
DII	6%	95%	0%	30%	6%
RNI	92%	67%	17%	23%	100%

tion is from 2005 and 2 citations are from 2010. The ‘publication age’ is (5 years + 10 years) = 15 years. We applied this algorithm to all papers of types article and letter from the period 1990–1994. No restrictions on the citation-windows or the minimum number of citations were imposed. In Table 6.12 we show the number of papers in both data sets Categories and Clusters that are selected by Redner’s algorithm, and also the overlap between this algorithm and our algorithms. In our opinion is what we call a ‘breakout’ identical to what Redner calls a ‘breakthrough’.

The robustness of Redner’s algorithm is tested in the same way that the other algorithms are tested. We show the results in Tables 6.13 and 6.14, and in Figure 6.7 and Figure 6.8.

Classification of breakouts

We classified the results of the algorithms across the two dimensions (1) the number of times cited by review papers and (2) the number of times cited by patents. Because the distribution of papers with a certain number of citations is skewed we focus on the Top 10% of the distribution. The top 10% percentile is subdivided into three ranges. In the analyses we use a grouping with the following four classes (1) ‘< Top 10%’ = [0.0-90.0)%, (2) ‘Top 10%’ = [90.0-95.0)%, (3) ‘Top 5%’ = [95.0-99.0)%, and (4) ‘Top 1%’ = [99.0-100.0]%. (see also Table 6.31). Table 6.7 shows that nearly 60% of the papers cited within 24 months are not in the Top 10% for both dimensions. In Table 6.32 the absolute numbers are presented. For all publications from 1990–1994 the share of papers outside the Top 10% simultaneously on both dimensions exceeds the 78% mark as is shown in the Tables 6.34 and 6.35. Based on this analysis we conclude that the algorithms distinguish papers along the following classification:

1. Breakthroughs: publications that are part of the scientific basis of Nobel Prize awarded discoveries.
2. Breakthrough by proxy’ publications that belong to the top 1% percentile on the basis of the number of citations from review publications and at the same time to the top 1% percentile of the number of citations from patents. These are the publication in the Top 1% row and at the same time in the Top 1% column.
3. Science-oriented breakthrough by proxy: publications that belong to the top 1% based on the number of citations from review publications but are not significantly cited from patents. These are the publications in the Top 1% row
4. Technology-oriented breakthrough by proxy: publications that not particularly highly by review publications but are in the top 1% based on citation from patents. These are the publications in the Top 1% column
5. Breakout: a publication identified by at least one of the implemented algorithms not belonging to one of the four types defined above.
6. Non-breakout: a paper not selected by any of the algorithms

In Table 6.6 we show the distribution of all 1990–1994 (articles & letters)

6.3 But is it a breakthrough?

Table 6.6: Distribution of all 1990-1994 (articles & letters) identified as a breakout paper, along the two dimensions 'cited by review papers' and 'cited by patents'. Based on citation counts from the moment of publication until 2015. (n=1,805,245)

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	17.53%	10.46%	8.25%	2.40%
Top 5%	15.70%	6.71%	3.41%	0.63%
Top 10%	9.94%	2.67%	1.15%	0.15%
< Top 10%	16.49%	3.39%	0.98%	0.13%

Table 6.7: Distribution of all 1990-1994 (articles & letters) cited within 24 months, along the two dimensions 'cited by review papers' and 'cited by patents'. Based on citation counts from the moment of publication until 2015. (n=408,292)

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	0.79%	0.13%	0.09%	0.03%
Top 5%	8.11%	0.41%	0.18%	0.03%
Top 10%	29.07%	0.52%	0.18%	0.02%
< Top 10%	59.44%	0.73%	0.23%	0.03%

Table 6.8: Effect of applying the algorithms to the complete dataset with papers (articles, letters) from 1990-1994 A '+' indicates that share of papers in this category increased, and a '-' that it decreased as result of the selection done by the algorithms.

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	+	+	+	+
Top 5%	-	+	+	+
Top 10%	-	+	+	+
< Top 10%	-	+	+	+

identified as a breakout paper, along the two dimensions ‘cited by review papers’ (rows) and ‘cited by patents’ (columns), and in Table 6.7 the information for publications from 1990-1994 identified as a breakout is presented. Comparing the information in these two tables shows the effect of applying the algorithms – the share of papers in the Top 10% percentiles increases. In Table 6.8 this effect is illustrated; a ‘+’ in a cell means that after applying the algorithms the concentration of the papers in this cell is higher than in comparable cell in the total dataset, and a ‘-’ means that the concentration is lower. The algorithms succeed in selecting at early stage papers that after 20+ years belong to the papers highly cited by review papers, and by patents. Detailed information for all publications from 1990-1994, the papers identified as a breakout and those not seen as a breakoout are shown in the Tables 6.32-6.40. For the papers from 1990-1994 that are not cited within the first 24 months after publication $\approx 85\%$ of the papers is outside the Top 10% percentiles. The data illustrates that no major error is made by not including papers that are not cited within 24 months, as even after 20+ years most these publications stay low cited, and do not fulfil the criteria to be seen as a *breakout* publication.

6.4 Discussion and overall findings

6.4.1 Number of citation at early stage as predictor for the total number of citations

Adams (2005) found that in general it is likely that a publication that is highly cited at early stage is also highly cited in the long run. In this study we find that a similar relation holds for citations a publication receives from review articles. Wang et al. (2013) built a model that adequately predicts the number of citation a publication will receive based on the number of citation in the first 5 years and conclude “Although additional variables combined with data mining could further enhance the demonstrated predictive power, an ultimate understanding of long-term impact will benefit from a mechanistic understanding of the factors that govern the research community’s response to a discovery.”

6.4.2 Citations from review articles

The breakthrough publications that form the basis of the four case studies (HIV/AIDS, Graphene, Ubiquitin, and Intron) received from review articles within 24 months minimal 4 citations, and until the beginning of 2016 at least 73. Of the papers in the validation test-set 32 belong to the top 1% percentile based on the citations from review papers received within the first 24 months after publication. These 32 publications in the top 1% percentile after 24 months are also in the top 1% in the beginning of 2016. We classify

Table 6.9: Key figures for Clusters and Categories

Dataset	Total number of papers	of which discovery-science papers (share)
Categories	253,558	91,775 (36%)
Clusters	214,827	87,061 (41%)
In Categories and in Clusters	160,943	62,767 (39%)
In Categories not in Clusters	92,785	29,008 (30%)
In Clusters not in Categories	53,965	24,924 (45%)

publications that are highly referenced by review articles – belong to the top 1% percentile – as potential breakthroughs.

6.4.3 The algorithms

Based on the concept of what characterizes a breakthrough we implemented and tested the following five breakout detection algorithms ‘Application-oriented Research Impact’ (ARI), ‘Discoverers Intra-group Impact’ (DII), ‘Researchers-Inflow Impact’ (RII), ‘Cross-Disciplinary Impact’ (CDI), and ‘Research-Niche Impact’ (RNI). The ARI, CDI, and RII algorithms were developed to identify ‘Charge’-type breakouts, and DII as well as RNI focus on the detection of ‘Challenge’-type breakouts. It appears, based on the numbers of breakout papers as presented in Table 6.2, that the algorithms make up three distinct groups. The algorithms RII and CDI that select the largest numbers of breakout papers populate one group. DII forms a second group by itself with an intermediate number of breakout papers. The third group is formed by the RNI and ARI algorithms that select the smallest numbers of papers.

The performance of ARI on the two datasets (Clusters and Categories) differs; it detects four times as many breakout papers in Categories as it does in Clusters. ARI searches for papers that are supposed to act as bridges between discovery-oriented science and application-oriented science. An analysis of the share of discovery-science papers in both datasets, their intersection, and differences are shown in Table 6.9.

Approximately 36% of the papers in the source data set are characterised as discovery-science oriented. Table 6.9 also shows that the share of discovery-science papers is above this average for Clusters and equal to this average or Categories. This finding in combination with the fact that the number of discovery-science papers is 5% lower in Clusters as compared to Categories makes that the behaviour of ARI seems to be counterintuitive. As the datasets are constructed from the same data source by a conceptually equivalent method - for each year and for each category or cluster the top 10% of papers most cited in the first 24 months after publication - the difference in performance of ARI on both datasets must be caused by a side effect of the applied selection method. Assuming that the probability for a paper

to be selected only depends on the citation characteristics of a paper the performance difference of ARI by a factor 4 is not caused by the difference in the number of discovery-science papers in both datasets as is illustrated in Table 6.9. The five factors that we believe play a role in this ‘ARI anomaly’ are:

1. The document selection process distributes the papers among 823 clusters (out of 865), and among 199 categories (out of 251). This results in an average of 106 discovery-science papers per cluster, and 461 per category;
2. For 60% of the 199 categories, the share of discovery-science papers is above the overall average of 36%; for the 823 clusters this share is equal to the overall average;
3. Discovery-science papers receive within 24 months on average 4.7 citations compared to 3.8 for application-science papers;
4. Papers can have more than one ‘subject category’ – on average 1.5 – assigned to them, but can be a member of only one document cluster. Therefore the same paper might occur multiple times for different categories during the selection of documents for Categories, thereby preventing other papers from being selected. The selection method we apply to build the Categories dataset may select the same paper for different categories. Hereby providing a bias towards highly cited papers to which multiple categories are assigned. In this way preventing other less cited papers to be selected;
5. On average fewer subject categories are assigned to discovery-science papers than to the more applied-science oriented papers.

In our opinion these factors in combination with the method of constructing the two data sets mean that higher cited discovery-science papers are preferred in the selection of papers for Categories, and thereby account for the higher performance of ARI on Categories than on Clusters.

The DII algorithm is the most sensitive of our five algorithms for thresholds imposed on the data. As ‘Charge’ breakouts are the more common variant, because there is no change in the theoretical framework or paradigm shift involved, it comes as no surprise that RII and CDI are the algorithms that select the most papers as a breakout. There is no ‘overall winner’ among the algorithms due to the fact that each is developed with a particular type of breakthrough paper in mind. But the conclusion that a breakout paper really presents a breakthrough must be based on information other than bibliographic information. The fact that the CDI algorithm behaves differently on both datasets is to be expected. This difference in behaviour is caused by the different definitions used in both datasets for the ‘discipline’ concept.

Comparing with Redner’s algorithm

The method proposed by Redner takes into account all citations, whereas the algorithms we developed focus on the citation dynamics of a paper within 24–36 months after publication. Except in the case of the DII algorithm - there is overlap between the results of our algorithms and Redner’s algorithm, some-

times a very small one; for details we refer to Table 6.12. Redner's algorithm selects 6,150 papers in Categories and 6,311 in Clusters; 5,907 of these papers belong to both datasets, 243 belong only to Categories, and 404 only to Clusters. From this we conclude that the performance of Redner's algorithm is largely independent of the dataset to which the algorithm is applied. This result is expected, as 'disciplines' are not addressed in Redner's algorithm, and therefore the differences are caused by differences in the contents of the datasets.

Redner's algorithm identifies in total 36 of the 60 papers in the test set. In Table 6.17 detailed information is provided for all 60 papers in the test set. In 12 cases the CDI algorithm when applied to the Clusters-set selected a paper, but not when applied to the Categories dataset; the coloured fields in Table 6.15 and Table 6.17 indicate these cases.

Overall findings:

- All our detection algorithms are able to identify breakout papers.
- Some of the breakout papers also stand out in citations given in patents and review papers, and are cited by social media sources;
- The analysis of the datasets Categories and Clusters shows the ability of each algorithm, regardless of the dataset, to select papers that are not selected by the other algorithms;
- The outcomes of the robustness calculations (Table 6.13, Table 6.14, Figure 6.8, and Figure 6.12) show that the algorithms CDI, RII, and RNI are the ones that - up to a threshold value of 32 citations - are almost unaffected by the value of the threshold;
- The five algorithms combined identified all 11 papers of WoS-type 'article' or 'letter' that were published in the period 1990-1994 and occur in Nature's 'Top-100 list of most cited papers ever';
- For five of the eight Nobel Prizes in Chemistry, Physics, and Physiology or Medicine for which scholarly work published between 1990 and 1994 forms the scientific basis, at least one of the founding papers was detected;
- In the test set 42 out of the 60 papers in the test set belong to the Top 2% of papers based on the number times cited in review articles, and 35 (58%) belong to the Top 0.5%;
- Half of the 60 papers in the test set are cited in patents, and therefore provide a link between science and technology;
- Except for one algorithm (DII) the selected papers are 'high' or 'very high' cited by review papers, they are cited in patents, and received citations within 24 months;
- Both ARI and CDI perform different when applied to Categories and Clusters; this is not the case for RII, DII, and RNI.
- The results produced by our detection algorithms, except DII, and the results produced by Redner's algorithm show overlap as indicated in Table 6.12. We notice that Redner's algorithm selects more publications.

6.4.4 Conclusions

The aim of this study is to develop general applicable algorithms to capture the dynamics of the diffusion of scholarly knowledge, and conclude at early stage if a paper should be considered a breakout. To decide which of the detected breakout papers really presents a breakthrough in science, we need additional information, particularly expert opinions. Our basic assumption is that the mechanisms responsible for a scientific discovery to evolve into a breakthrough already provide characteristic signals in the bibliographic information at an early stage. By focusing on the first 24–36 months after publication of a paper we ignore ‘sleeping beauties’ (van Raan, 2004, 2015). We also did not address the situation in which the citation profile of a paper at an early stage give the impression that it presents a ‘breakthrough’, that at a later stage turns out not to be the case. An example is (Fleischmann and Pons, 1989) in which the authors claim the existence of nuclear fusion at room temperature ‘cold fusion’. It was almost immediately criticized, and it was concluded that “... According to our calculations, the experimentally measured excess heat can be accounted for fully by this chemical reaction ...” (Dmitriyeva et al., 2012).

The retraction of scientific publications is increasing, and the number of retracted papers in Medline (: U.S. National Library of Medicine® (NLM) premier bibliographic database) reached the 1% level in 2006 (Cokol et al., 2008). Retracted publications however do not vanish from the scientific-knowledge base and are still cited even after their retraction, and in only 8% of the citations the retraction is mentioned (van Noorden, 2011). Retracted articles live on in personal libraries and on the Internet (Davis, 2012). Publications that at some point in time will be retracted are therefore in general present in the data selected as input to the algorithms.

Within the context described above, this study addresses the following three objectives:

1. Construction of general-purpose algorithms to detect breakout papers;
2. Comparison of the algorithms on their ability to detect breakout papers;
3. Determination of the effectiveness of the algorithms in identifying breakout papers.

To do so we constructed and applied five different citation-based algorithms: ‘Application-oriented Research Impact’ (ARI), ‘Discoverers Intra-group Impact’ (DII), ‘Researchers Inflow Impact’ (RII), ‘Cross-Disciplinary Impact’ (CDI), and ‘Research Niche Impact’ (RNI). These five algorithms can be divided into three groups based on the breakout-detection specificity (recall rate). Group 1 consists of the CDI and RII algorithms. For these algorithms the recall rate increases with increasing threshold values, and reaching above a certain threshold value (64 for RII, and 128 for CDI) the situation where the algorithm selects all remaining papers. The second group consists of ARI and RNI. These algorithms also show an increasing recall rate, but above a certain threshold value (32 for ARI, and 128 for RNI) they break down and fail to select any documents. DII forms a group by itself as the recall rate continuously decreases with increasing threshold values; detailed information can be found

in Tables 6.24 and 6.25, and is illustrated with Figure 6.10 and Figure 6.11.

The detection algorithm developed by Redner (2005) can be considered a high-performance algorithm and therefore falls in group 1 together with our CDI and RII algorithms. For threshold values from 256 citations and above, these algorithms select the same documents as breakout papers.

For each algorithm we calculated (Table 6.26) the probability to select a paper as breakout paper in a dataset (Categories or Clusters) when no thresholds are applied. Using these probabilities we calculated for each algorithm the number of breakout papers that is to be expected when different thresholds are applied on the datasets. We also compared our results with the results obtained after applying Redner's algorithm. The conclusion is that, except for DII, all algorithms select more papers than expected which means that these algorithms select 'genuine' breakout papers and are less sensitive to changes in the thresholds; details can be found in Tables 6.26–6.30. The behaviour of the DII algorithm is different as it focuses on work where the discovery involves a paradigm-shift that starts within a small group of researchers; the core group. Given the short measuring period after publication, the probability for such a paper to get cited by authors outside of the core group is limited; therefore the number of papers selected by DII shows a sharp decrease for larger threshold values.

The remarkable observation that RNI, our 'lowest-output' algorithm, has such a high hit rate in terms of selecting breakout papers raises the question 'How effective is the RNI algorithm in detecting breakthroughs?' Further in-depth research is needed to answer this question as no definitive answer can be given on the basis of the available bibliographic information.

As there is no objective measure to identify a breakthrough at early stage, we used the impact of the breakout papers as a measure. The findings in this study suggest that the early detection algorithms that we developed provide added value for tracing breakout papers. We tried to validate the outcomes of the algorithms against other information sources like Nobel Prize awards, Nature's 'Top-100 list of papers most cited ever', and citations from patents, review papers and social media. The conclusion of this validation is that part of the identified breakout papers are considered to be important for developments in science, and some of these papers might indeed be considered a breakthrough. We therefore conclude that our algorithms are able to detect important scientific papers at an early stage and provide opportunities to reap the benefits of new discoveries sooner; compared to simple citation counts only; we believe that our algorithms have considerable added value in identifying breakthroughs in science at early stage.

In our opinion we prove our hypothesis "It is possible to design, develop, implement, and test an analytical framework and measurement model as a general-purpose tool using bibliographic information for early detection of breakthroughs in science?" to be true. On-going and future research will also include further refinement of the current algorithms, investigation of the similarities and differences between the algorithms, and the construction of new algorithms for early stage identification of breakout papers.

6.5 Acknowledgments

We thank Professor Redner for answering questions on details of his algorithm, and our colleague Ludo Waltman for his help in applying the CWTS publication-clustering method to produce the dataset Clusters. We also like to thank David Pendlebury for providing insight into the method used by Thomson Reuters to identify *Nobel prize class* papers.

Appendix 6.A Supplementary material

This supporting information is organized in nine subsections:

6.A.1 The datasets: Categories and Clusters

Describes the two datasets Categories and Clusters, and shows the distribution of the number of papers versus the number of citations received in 24 months for all papers (articles, letters) published in 1990–1994 and the papers in the Categories and Clusters dataset.

6.A.2 Effect of the CWTS document-clustering method on ‘multidisciplinarity’

This subsection illustrates the effect of the CWTS document-clustering algorithm, which is used to construct the Clusters dataset, on the ‘multidisciplinarity’ measure.

6.A.3 Overlap with Redner’s algorithm

The outcomes of Redner’s algorithm are compared with the outcomes of the five algorithms that make up the analytical framework.

6.A.4 Robustness of the algorithms

This subsection shows the robustness of the five algorithms (ARI, CDI, RII, DII, RNI) developed in this study and Redner’s algorithm when applied to the datasets for Clusters and Categories. It presents the overlap in the outcomes of the five algorithms and those produced by Redner’s algorithm.

6.A.5 Performance of the algorithms

The performances of the implemented algorithms was tested by looking at papers from 1990–1994 considered fundamental to Nobel Prize awarded research, or papers that appear in Nature’s Top 100 list of papers most cited ever. In addition, a test set of 60 papers was used to examine citations from review papers, patents, and social media (2012–2014).

6.A.6 Key figures of the papers in the test set

In this chapter the outcomes of applying the algorithms to the test set of 60 papers are further broken down.

6.A.7 Breakout detection specificity (recall rate) of the algorithms

The recall rate for the algorithms is presented for various threshold values.

6.A.8 Detection probability and expected breakout recall rate of the algorithms

In this subsection the performance of the algorithms is compared with what might be expected.

6.A.9 Breakout classification

In this section the ‘breakout character’ of the results after applying the algorithms is presented.

Table 6.10: Key figures for the two datasets used

Dataset	Number of papers
Categories	253,558
Clusters	214,827
In Categories and in Clusters	160,943
In Categories not in Clusters	92,785
In Clusters not in Categories	53,965

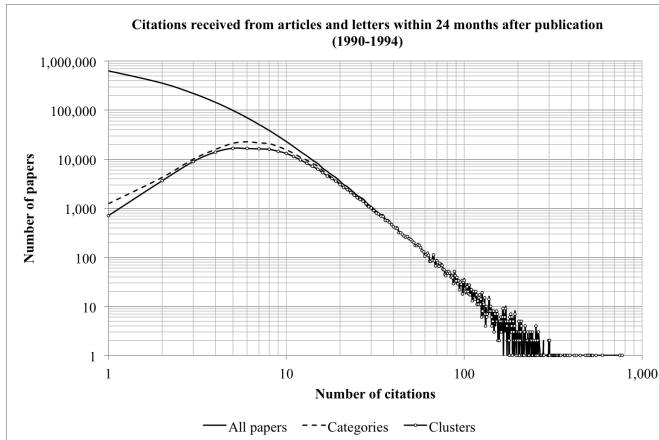


Figure 6.3: Distribution of the number of citations within 24 months for all papers (articles and letters published in the period 1990–1994), and papers in the datasets Categories, and Clusters

6.A.1 The datasets: Categories and Clusters

In the studies we used two datasets – Categories and Clusters. In this section the key figures for these datasets are presented. Table 6.10 shows for the datasets Categories and Clusters the numbers of papers, and also the overlap and the differences. Figure 6.3 shows for all papers (articles and letters) covered in the WoS published in the period 1990–1994, the papers in Categories, and those in Clusters the distribution of the number of papers and the number of citations received within the first 24 months after publication. The effect of the selection method used to construct the datasets Categories and Clusters, which is explained in the methods section of the main text, is clearly visible in this figure.

The cumulative citation distributions for papers cited by review papers is shown in Figure 6.4 and for papers cited in patents in Figure 6.5. Shown is the fact that for these those parameters ‘Breakout papers’ obtain on average higher more citations from review papers and from patents than ‘Not-breakout papers’. Included are in papers (articles and letters) published in the period 1990–1994, and covered by the WoS.

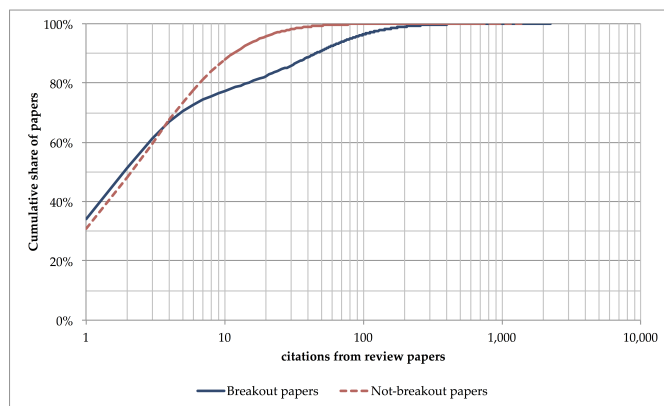


Figure 6.4: Cumulative distribution for number of papers vs. number of times cited by review papers. Included are all publication (articles, letters) in the WoS published in 1990–1994

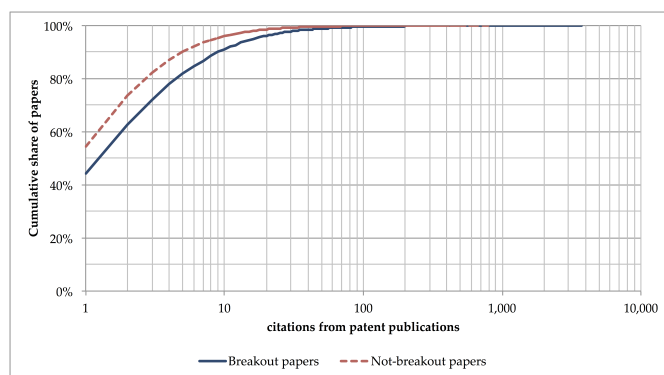


Figure 6.5: Cumulative distribution for number of papers vs. number of times cited by patents. Included are all publication (articles, letters) in the WoS published in 1990–1994

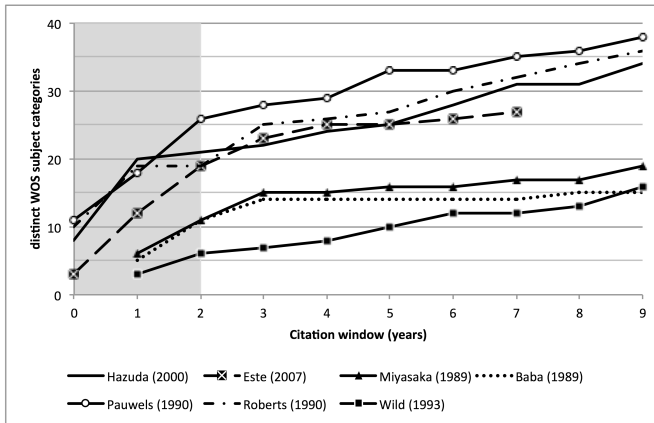


Figure 6.6: Cross Disciplinary Impact for a sample of seven papers using as measure the number of different ‘WoS subject categories’ of the citing papers at a certain point in time

6.A.2 Effect of the CWTs document-clustering method on ‘multidisciplinarity’

This section illustrates the effect the CWTs document-clustering, which is used to construct the Clusters dataset, has on ‘multidisciplinarity’⁹⁸. Figure 6.6 and Figure 6.7 show the differences for the seven papers that were used to determine the parameters for the ‘Cross Disciplinary Impact’ algorithm (CDI); the differences are particularly noticeable in the greyish areas that indicate the period of particular interest in this study. The Cross Disciplinary Impact (CDI) algorithm comes therefore in two flavours, one for Categories (CDI_{sc}) and the other for CWTs document clusters (CDI_{dc}). The X-axis in these figures shows the number of years after the publication, and the Y-axis the multidisciplinarity of the documents citing the publications. The bibliographic information for the seven papers is presented in Table 6.11.

6.A.3 Overlap with Redner’s algorithm

Redner introduced an algorithm to identify — what he calls — breakthrough papers (Redner, 2005). Redner took an approach different from the one we took in constructing our algorithms. It is therefore interesting to compare the outcomes of our algorithms with those of Redner’s algorithm. Table 6.12 shows the results.

⁹⁸Multidisciplinarity is defined as the number of different categories or clusters. Categories and clusters are the basic categorisation dimension in the two datasets

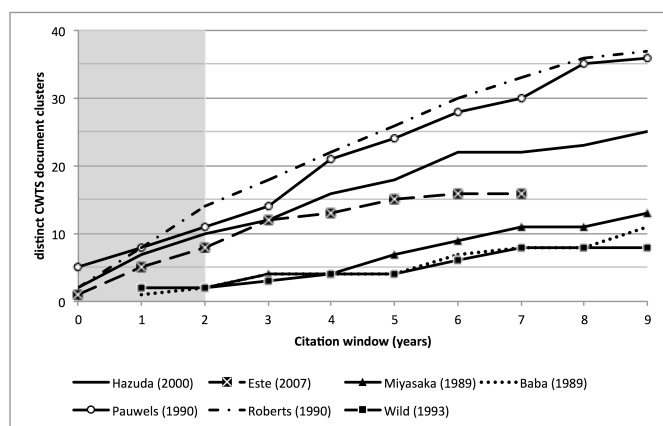


Figure 6.7: Cross Disciplinary Impact for a sample of seven papers using as measure the number of different ‘CWTs paper clusters’ of the citing papers at a certain point in time.

Table 6.11: Bibliographic information for the papers mentioned in Figure 6.6 and Figure 6.7

- Baba, M., Tanaka, H., De Clercq, E., Pauwels, R., Balzarini, J., Schols, D., Nakashima, H., Perno, C. F., Walker, R. T., and Miyasaka, T. (1989). Highly specific-inhibition of human immunodeficiency virus type-1 by a novel 6-substituted acyclouridine derivative. *Biochemical and Biophysical Research Communications*, 165(3):1375–1381.
- Esté, J. A. and Telenti, A. (2007). HIV entry inhibitors. *The Lancet*, 370(9581):8–88.
- Hazuda, D. J., Felock, P., Witmer, M., Wolfe, A., Stillmock, K., Grobler, J. A., Espeseth, A., Gabryelski, L., Schleif, W., Blau, C., and Miller, M. D. (2000). Inhibitors of strand transfer that prevent integration and inhibit HIV-1 replication in cells. *Science*, 287(5453):646–650.
- Miyasaka, T., Tanaka, H., Baba, M., Hayakawa, H., Walker, R. T., Balzarini, J., and De Clercq, E. (1989). A novel lead for specific anti-HIV-1 agents: 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine. *Journal of Medicinal Chemistry*, 32(12):2507–2509.
- Pauwels, R., Andries, K., Desmyter, J., Schols, D., Kukla, M. J., Breslin, H. J., Raeymaeckers, A., Gelder, J. V., Woestenborghs, R., Heykants, J., Schellekens, K., Janssen, M. A. C., De Clercq, E., and Janssen, P. A. J. (1990). Potent and selective inhibition of HIV-1 replication in vitro by a novel series of TIBO derivatives. *Nature*, 343(6257):470–474.
- Roberts, N., Martin, J., Kinchington, D., Broadhurst, A., Craig, J., Duncan, I., Galpin, S., Handa, B., Kay, J., Krohn, A., Lambert, R., Merrett, J., Mills, J., Parkes, K., Redshaw, S., Ritchie, A., Taylor, D., Thomas, G., and Machin, P. (1990). Rational design of peptide-based HIV proteinase inhibitors. *Science*, 248(4953):358–361.
- Wild, C., Greenwell, T., and Matthews, T. (1993). A synthetic peptide from HIV-1 gp41 is a potent inhibitor of virus-mediated cell–cell fusion. *AIDS Research and Human Retroviruses*, 9(11):1051–1053.

Table 6.12: Overlap of the results of Redner's algorithm and the five algorithms (ARI, DII, RII, CDI, RNI) (no threshold applied)

Algorithms	Categories	Clusters
	Number of papers marked as breakout	
Redner	6,150	6,311
Redner $\cap$ ARI	8	11
Redner $\cap$ CDI	943	3,210
Redner $\cap$ RII	2,119	2,108
Redner $\cap$ DII	0	0
Redner $\cap$ RNI	13	6

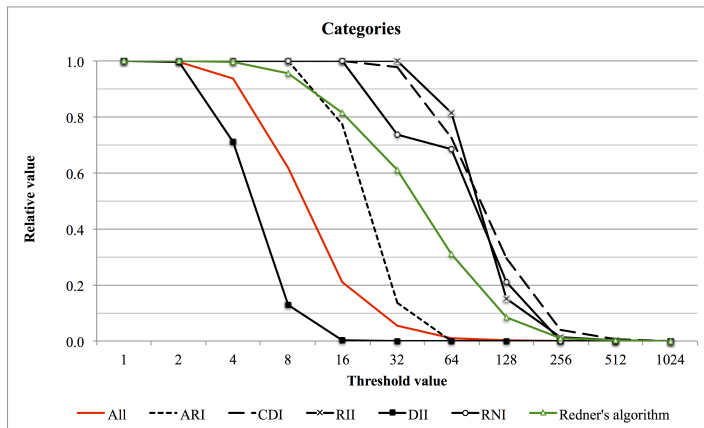


Figure 6.8: Categories: normalized number of papers as a function of the applied threshold. For each variable the number of papers for the threshold value 1 is taken as reference.

6.A.4 Robustness of the algorithms

Robustness in this context is the ability of an algorithm to produce the same results irrespective of the different thresholds applied to the data. This subsection shows the robustness of the five algorithms (ARI, CDI, RII, DII, RNI) developed in this study and Redner's algorithm when applied to the datasets for Clusters and Categories. It presents the overlap in the outcomes of the five algorithms and those produced by Redner's algorithm. Tables 6.13 and 6.14 contain detailed information for Categories and Clusters on the outcomes for the five algorithms ARI, DII, RII, CDI, RNI, and for the algorithm introduced by Redner (Redner, 2005) when thresholds on the number of citations a paper received within 24 months are applied. In Figure 6.7 and 6.8 we present the number of papers as a share of the number of papers for the situation when no threshold is applied (≥ 1); this is done for the datasets as a whole ('All') and for the outcomes of the various algorithms.

Table 6.13: Categories: number of papers in the Categories dataset, and in the various result sets in relation to the applied threshold

	Categories dataset	ARI	CDI _{sc}	RII	DII	RNI	Redner's algo- rithm
Threshold (Number of times cited within 24 months)							
≥ 1	253,558	264	1,276	3,543	577	19	6,150
≥ 2	252,316	264	1,276	3,543	576	19	6,150
≥ 4	238,009	264	1,276	3,543	410	19	6,133
≥ 8	156,765	264	1,276	3,543	74	19	5,883
≥ 16	53,014	204	1,275	3,543	2	19	5,017
≥ 32	13,583	36	1,246	3,543	0	14	3,748
≥ 64	2,996	0	924	2,891	0	13	1,915
≥ 128	539	0	375	539	0	4	511
≥ 256	55	0	53	55	0	0	55
≥ 512	7	0	7	7	0	0	7
≥ 1024	0	0	0	0	0	0	0

Table 6.14: Clusters: number of papers in the dataset and in the various result sets in relation to the applied threshold

	Categories dataset	ARI	CDI _{dc}	RII	DII	RNI	Redner's algo- rithm
Threshold (Number of times cited within 24 months)							
≥ 1	214,827	60	13,477	3,501	674	8	6,311
≥ 2	214,119	60	13,477	3,501	673	8	6,311
≥ 4	201,514	60	13,477	3,501	483	8	6,299
≥ 8	137,969	60	13,451	3,501	74	8	5,839
≥ 16	51,666	60	12,096	3,501	2	8	5,179
≥ 32	13,369	56	6,930	3,501	0	7	3,748
≥ 64	2,926	0	2,320	2,857	0	7	1,904
≥ 128	534	0	486	534	0	4	508
≥ 256	54	0	54	53	0	0	54
≥ 512	7	0	7	7	0	0	7
≥ 1024	0	0	0	0	0	0	0

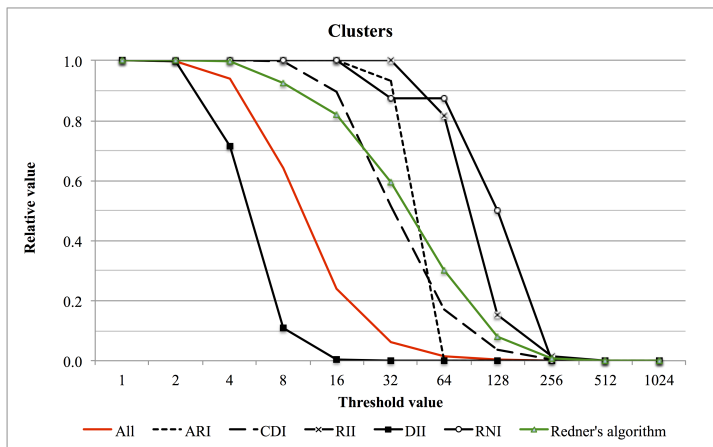


Figure 6.9: Clusters: normalized number of papers as a function of the applied threshold. For each variable the number of papers for the threshold value 1 is taken as reference.

6.A.5 Performance of the algorithms

The performances of the implemented algorithms was tested by looking at papers from 1990–1994 considered fundamental to Nobel Prize awarded research, or papers that appear in Nature’s Top 100 list of papers most cited ever. In addition, a test set of 60 papers was used to examine citations from review papers, patents, and social media (2012–2014). The five algorithms (ARI, DII, RII, CDI, RNI) were tested against (1) Nature’s Top-100 list of papers most cited ever, (2) Nobel prizes based on scholarly work from the period 1990–1994, (3) the test set of 60 papers.

Nature’s Top-100 list of papers most cited ever

Table 6.15 shows the results of the test of the paper in the test set against Nature’s Top-100 list of papers most cited ever’. The bibliographic information for these papers is presented in Table 6.19.

Table 6.15: Results for the 1990–1994 papers in the top 100 most cited papers of all times (coloured marks indicate that only the CDI_{dc} algorithm recognised the paper in combination with the Clusters dataset)

Paper	Total citations received	Algorithm					Comment
		ARI	CDI	RII	DII	RNI	
Altschul et al (1990)	37,635		X	X			1990 paper most cited by review papers
Becke (1993)	48,292		X	X			1993 paper that belongs to the most cited by review papers within the top 0.5% most cited papers
Blöchl (1994)	14,004		X				Top 0.5% most cited papers
Iijima (1991)	20,581		X	X			1991 paper most cited by review papers
Jones (1991)	12,479		X	X			Top 0.5% most cited papers
Kraulis (1991)	13,009		X	X			1991 paper that belongs to the most cited by review papers within the top 0.5% most cited papers
Laskowski et al. (1993)	15,238						Not included, paper has document type 'Software Review'
Moncada et al. (1991)	13,670						Not included, paper has document type 'Review'
O'Regan and Grätzel (1991)	13,169		X				1991 paper that belongs to the most cited by review papers within the top 0.5% most cited papers
Perdew and Wang (1992)	12,796		X				Top 0.5% most cited papers
Sheldrick. (1990)	18,165		X	X			Top 0.5% most cited papers
Thompson et al. (1994)	39,278		X	X			1994 paper that belongs to the most cited by review papers within the top 0.5% most cited papers
Ware and Sherbourne (1992)	13,389		X	X			1992 paper most cited by review papers

Scholarly papers from the period 1990–1994 considered as the foundation of Nobel Prize awarded research

We applied the algorithms to the scholarly papers from the period 1990–1994 on which Nobel Prize prize-winning research in the fields Chemistry, Physics, and Physiology or Medicine is based. Table 6.16 shows the results.

Table 6.16: Nobel Prizes based on scholarly work published in the period 1990–1994; indication is given of whether a paper is recognised and by which algorithm

Nobel Prize Category (year)	Relevant Scholarly papers according to the Nobel Prize Committee for the Nobel Prizes in Chemistry, Physics, and Physiology or Medicine
Chemistry (1997)	Boyer, P.D. (1993). The binding change mechanism for ATP synthase - Some probabilities and possibilities. <i>Biochimica et Biophysica Acta</i> 1140:215–250. (paper not in dataset) Abrahams. J.P., Leslie. A.G., Lutter. R., and Walker J.E. (1994). Structure at 2.8 Å resolution of F₁-ATPase from bovine heart mitochondria. <i>Nature</i>. 370:621–628. (recognised by RII and CDI) Skou. J.C., and Esmann. M. (1992). The Na. K-ATPase. <i>Journal of Bioenergetics and Biomembranes</i>. 24:249–261. (paper not in dataset)
Physics (2001)	C. Monroe, W. Swann, H. Robinson and C. Wieman.(1990). <i>Phys. Rev. Lett.</i> 65:1571 (not recognised) C. Monroe. E. Cornell and C. E. Wieman. page 361 in Proceedings of the International School of Physics 'Enrico Fermi'. Course CVXIII. Laser Manipulation of Atoms and Ions (North Holland. Amsterdam. 1992) (paper not in dataset) W. Ketterle, K. B. Davis, M. A. Joffe, A. Martin and D. E. Pritchard.(1993) <i>Phys. Rev. Lett.</i> 70:2253 (not recognised)
Chemistry (2003)	Preston. G. M., and Agre. P. (1991). Isolation of the cDNA for erythrocyte integral membrane protein of 28 kilodaltons: member of an ancient channel family. <i>Proc Natl Acad Sci USA</i> 88. 11110–11114. (recognised by CDI on Clusters) Preston. G. M.. Carroll. T. P.. Guggino. W. B.. and Agre. P. (1992). Appearance of water channels in <i>Xenopus</i> oocytes expressing red cell CHIP28 protein. <i>Science</i> 256. 385–387. (recognised by CDI on Clusters) Zeidel. M. L.. Ambudkar. S. V.. Smith. B. L.. and Agre. P. (1992). Reconstitution of functional water channels in liposomes containing purified red cell CHIP28 protein. <i>Biochemistry</i>. 31:7436–7440. (not recognised)
Physiology or Medicine (2004)	Buck. L. and Axel. R. (1991) <i>Cell</i>. vol. 65:175–187. (recognised by RII and CDI)

6.A Supplementary material

Table 6.16 (continued from previous page)

Nobel Prize Category (year)	Relevant Scholarly papers according to the Nobel Prize Committee for the Nobel Prizes in Chemistry, Physics, and Physiology or Medicine
Physics (2006)	Mather, J.C. et al. (1990) <i>Astrophys. J.</i> (Letter) 354:37 (recognised by RII) Mather, J.C. (1994). <i>Astrophys. J.</i> 420. 440 (recognised by RII) Smoot, G. et al. (1990) <i>Astrophys. J.</i> 360. 685 (!volume number in Nobel prize document is incorrect (380)). (not recognized) Smoot, G. et al. (1992) <i>Astrophys. J.</i> (Letters) 396. 1 (recognised by RII and CDI)
Chemistry (2012)	Cotecchia, S., Exum, S., Caron, MG. and Lefkowitz RJ (1990) Regions of the alpha 1-adrenergic receptor involved in coupling to phosphatidylinositol hydrolysis and enhanced sensitivity of biological function. <i>Proc Natl Acad Sci USA.</i> 87:2896-2900. (not recognised)
Physics (2014)	Akasaki H., Amano, K., Itoh, N., Koide and K. Manabe.(1992). <i>Int. Phys. Conf. Ser.</i> 129:851. (paper not in dataset) S. Nakamura, M. Senoh and T. Mukai. <i>Jpn. J. Appl. Phys.</i>. 32. L8 (1993) (not recognised) S. Nakamura et al., (1993). <i>J. Appl. Phys.</i> 74:3911 . (paper not in dataset) S. Nakamura, T. Mukai, and M. Senoh. (1994). <i>Appl. Phys. Lett.</i> 64:1687. (recognised by RII) S. Nakamura, (1991). <i>Jpn. J. Appl. Phys.</i>. 30:L1705 (not recognised) S. Nakamura, M. Senoh and T. Mukai. (1991) <i>Jpn. J. Appl. Phys.</i>. 30:L1998. (not recognised)
Chemistry (2014)	Hell, S.W. and Wichman, J. (1994). Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion-microscopy. <i>Opt.Lett.</i>. 19:780-782. (paper not in dataset)

Test set

After applying the five algorithms to the papers in the datasets Categories en Clusters, we selected for each combination of algorithm and dataset the papers in the top 10 based on times cited within 24 months after publication. The result is a set of 60 unique papers, and we show the key indicators for these papers in Table 6.17. The bibliographic information for these papers is contained in Table 6.19. We applied not only our five algorithms but also Redner's algorithms. In this table we also show if a paper is identified, and if so by which algorithm. For the papers in the test set we checked whether they occur on Nature's 'Top-100 list of most cited papers', whether they are cited in social media papers from the period 2012-2014, the number of times these papers are cited as non-patent literature in patents, and the times they are cited by review papers; the results are shown in Table 6.18.

Table 6.17: Papers in the test set, selected on the basis of belonging to the top 10 most cited papers within 24 months for each of the algorithms (an 'X' in a cell indicates that the paper is recognised as a breakout paper by the corresponding algorithm; coloured marks indicate that only the CDI_{dc} algorithm recognised the paper in combination with the Clusters dataset)

Paper	Dataset	Times cited ^a			Detecting algorithm					
		(total)	(≤ 12 months)	(≤ 24 months)	ARI	CDI	RII	DII	RNI	REDNER
Adeva et al. (1990)	Clusters	471	24	50	X					
Ahmet (1991)	Clusters	452	20	50	X					
Altschul et al. (1990)	Clusters + Categories	37,635	21	161		X	X			X
Angell (1991)	Clusters + Categories	1,146	3	39	X	X				X
'Anonymous' (1991)	Categories	422	19	42	X					
'Anonymous' ^b (1992)	Clusters + Categories	1,539	64	193		X	X		X	X
Bailey (1994)	Clusters + Categories	17,178	140	479		X	X			X
Becke (1993)	Clusters + Categories	48,292	7	75		X	X			X
Blöchl (1994)	Clusters	14,004	9	24		X				X
Brown et al. (1990)	Clusters + Categories	1,207	45	169		X	X		X	X
Coussot et al. (1993)	Categories	94	1	6				X		
Delerue et al. (1993)	Clusters + Categories	639	6	32	X					X
Dixon et al. (1994)	Clusters + Categories	1,095	7	51	X	X				X
Duschek et al. (1990)	Categories	91	1	4				X		
Eldeiry et al. (1993)	Categories	6,078	171	597		X	X			X
Fang et al. (1994)	Clusters	100	2	8				X		
Fischman et al. (1994)	Clusters + Categories	2,503	42	199		X	X		X	
Frabetti et al. (1992)	Clusters + Categories	110	11	19				X		
Frostig et al. (1992)	Categories	187	2	3				X		

6.A Supplementary material

(Table 6.17 continued from previous page)

Paper	Dataset	Times cited ^a			Detecting algorithm					
		(total)	(≤ 12 months)	(≤ 24 months)	ARI	CDI	RII	DII	RNI	REDNER
Gasser et al. (1993)	Clusters	211	2	7				X		
Gavrieli et al. (1992)	Categories	7,741	19	154		X	X			X
Hafner et al. (1992)	Categories	275	3	8				X		
Hodges et al. (1992)	Categories	856	3	17	X					X
Huntington (1993)	Clusters + Categories	1,042	33	88		X	X		X	X
Iijima (1991)	Clusters + Categories	20,581	55	182		X	X			X
Inoue et al. (1990)	Clusters + Categories	162	3	8				X		
Jones et al. (1991)	Clusters + Categories	12,479	13	65		X	X			X
Keller et al. (1993)	Clusters + Categories	591	10	33	X	X				X
Kessler et al. (1994)	Categories	6,302	44	193		X	X			X
Kiranoudis et al. (1993)	Clusters + Categories	128	3	6				X		
Koudstaal et al. (1993)	Categories	879	27	68			X		X	X
Kraulis (1991)	Clusters + Categories	13,009	39	234		X	X			X
Kubota et al. (1992)	Clusters	547	15	47	X					X
Kumbharkhane et al. (1991)	Clusters	101	3	4				X		
Larsson et al. (1994)	Clusters	119	2	3				X		
Lauck et al. (1990)	Clusters	104	2	7				X		
Lebel et al. (1990)	Categories	186	1	6				X		
Manabe et al. (1993)	Categories	438	9	24	X					
Mccormick and Pape (1990)	Clusters	641	9	42	X	X				X
Middlemas et al. (1991)	Clusters	582	14	43	X	X				X
Miller et al. (1992)	Clusters	373	37	76		X	X		X	
Ohara et al. (1993)	Categories	405	10	45	X					
O'Regan and Grätzel (1991)	Clusters	13,169	12	49		X				X

(Table 6.17 continued from previous page)

Paper	Dataset	Times cited ^a			Detecting algorithm					
		(total)	(≤ 12 months)	(≤ 24 months)	ARI	CDI	RII	DII	RNI	REDNER
Pause and Sonenberg (1992)	Categories	413	8	32	X					
Probstfield (1991)	Clusters + Categories	2,260	51	164		X	X		X	X
Riggs et al. (1990)	Categories	567	26	84		X	X		X	X
Rudman et al. (1990)	Categories	764	32	78		X	X		X	X
Ryder (1993)	Clusters	39	26	30					X	
Sakai et al. (1990)	Clusters	127	2	6				X		
Shamoon et al. (1993)	Clusters + Categories	9,957	141	449		X	X			X
Sheldrick (1990)	Clusters + Categories	18,165	13	84		X	X			X
Stampfer et al. (1991)	Categories	1,345	29	87			X		X	
Stave et al. (1990)	Categories	96	2	10				X		X
Svoboda et al. (1994)	Clusters + Categories	101	5	5				X		
Swedberg et al. (1992)	Clusters + Categories	518	36	115		X	X		X	X
Thompson et al. (1994)	Clusters + Categories	39,278	57	326		X	X			X
Ware and Sherbourne (1992)	Clusters + Categories	13,389	23	76		X	X			X
Whitesides and Laibinis (1990)	Clusters + Categories	806	7	50	X	X				
Williams et al. (1990)	Categories	7,441	24	141		X	X			X
Zhang et al. (1994)	Categories	6,058	84	342		X	X			X
Total number of papers from the test set recognized					15	32	26	16	12	36

^a Only citations from papers of the types 'article' or 'letter' are taken into account

^b According to the on-line version of Thomson Reuters Web of Science this paper is never cited, whereas in the CWTs's in house version of the database it is cited 1,539 times by papers of the type 'article' or 'letter'

Table 6.18: Validation of the papers mentioned in Table 6.16

Paper	Position in Nature's list of Top-100 most cited papers	Times cited by review papers	Cited by patent families (PATSTAT spring 2014 edition)	Cited by social media papers (2012-2014)	Notes
Adeva et al. (1990)		38			
Ahmet (1991)		37			
Altschul et al. (1990)	12	1,155	2,412	1 Blog post	BLAST software
Angell (1991)		72			
"Anonymous" (1992)		34	1		
"Anonymous" (1991)		350			
Bailey (1994)		149	18		CCP4 software
Becke (1993)	8	1,664	6	1 Twitter post, 1 Facebook post, 1 News message, 1 Blog post	
Blöchl (1994)	85	162	1		
Brown et al. (1990)		385	16		
Coussot et al. (1993)		4			
Delerue et al. (1993)		37	1		
Dixon et al. (1994)		98			
Duschek et al. (1990)		9			
Eldeiry et al. (1993)		942	6		
Fang et al. (1994)		1			
Fischman et al. (1994)		378	24	1 Twitter post	
Frabetti et al. (1992)		7			

Table 6.18 (continued from previous page)

Paper	Position in Nature's list of Top-100 most cited papers	Times cited by review papers	Cited by patent families (PATSTAT spring 2014 edition)	Cited by social media papers (2012-2014)	Notes
Frostig et al. (1992)		2			
Gasser et al. (1993)		11			
Gavrieli et al. (1992)		457	34		
Hafner et al. (1992)		38			
Hodges et al. (1992)		150			1 of two most cited papers in Brain 1985-1994)
Huntington (1993)		97		8 Twitter posts	
Iijima (1991)	36	1,509	46	69 Twitter posts, 1 Facebook post, 1 Blog post	
Inoue et al. (1990)		6	2		
Jones et al. (1991)	95	180	15	1 Blog post	
Keller et al. (1993)		100	15		
Kessler et al. (1994)		871	1		
Kiranoudis et al. (1993)		3			
Koudstaal et al. (1993)		300			
Kraulis (1991)	82	832	6		
Kubota et al. (1992)		18			
Kumbharkhane et al. (1991)					
Larsson et al. (1994)		5			
Lauck et al. (1990)		16			
Lebel et al. (1990)		15			

6.A Supplementary material

Table 6.18 (continued from previous page)

Paper	Position in Nature's list of Top-100 most cited papers	Times cited by review papers	Cited by patent families (PATSTAT spring 2014 edition)	Cited by social media papers (2012-2014)	Notes
Manabe et al. (1993)		35			
Mccormick and Pape (1990)		118			
Middlemas et al. (1991)		91	6		
Miller et al. (1992)		48			
Ohara et al. (1993)		736	1		
O'Regan and Grätzel (1991)	90	141	264	4 Blog posts, 1 Facebook post	
Pause and Sonenberg (1992)		58	2		
Probstfield (1991)		684			
Riggs et al. (1990)		188	2	2 Twitter posts	
Rudman et al. (1990)		282	6	2 Blog posts, 1 Facebook post, 2 Twitter posts, 1 Google+ post	
Ryder (1993)		1			
Sakai et al. (1990)		4	1		
Shamoon et al. (1993)		2133	16		
Sheldrick (1990)	47	178			
Stampfer et al. (1991)		275	1		
Stave et al. (1990)		6			
Svoboda et al. (1994)		7			
Swedberg et al. (1992)		197	2		
Thompson et al. (1994)	10	1154	475		

Table 6.18 (continued from previous page)

Paper	Position in Nature's list of Top-100 most cited papers	Times cited by review papers	Cited by patent families (PATSTAT spring 2014 edition)	Cited by social media papers (2012-2014)	Notes
Ware and Sherbourne (1992)	72	1,102	5		
Whitesides and Laibinis (1990)		80	5		1 of the top-25 most cited Langmuir papers in the period 1985-2009
Williams et al. (1990)		379	33		
Zhang et al. (1994)		1,607	83	2 Blog posts, 1 News item	

Table 6.19: Bibliographic information for the papers mentioned in Tables 6.15, 6.16, and 6.17

- Adeva, B., et al. (1990). The Construction of the L3 Experiment. *Nuclear Instruments & Methods in Physics Research Section A-Accelerators Spectrometers Detectors and Associated Equipment*, 289(1-2):35-102.
- Ahmet, K. (1991). The Opal Detector at LEP. *Nuclear Instruments & Methods in Physics Research Section A-Accelerators Spectrometers Detectors and Associated Equipment*, 305(2):275-319.
- Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403-410.
- Angell, C. (1991). Relaxation in Liquids, Polymers and Plastic Crystals - Strong Fragile Patterns and Problems. *Journal of Non-Crystalline Solids*, 131(1):13-31.
- "Anonymous" (1991). The Delphi Detector at LEP. *Nuclear Instruments & Methods in Physics Research Section A-Accelerators Spectrometers Detectors and Associated Equipment*, 303(2):233-276.
- "Anonymous" (1992). Systemic Treatment of Early Breast-Cancer by Hormonal, Cytotoxic, or Immune Therapy - 133 Randomized Trials Involving 31000 Recurrences and 24000 Deaths Among 75000 Women .1. *Lancet*, 339(8784):1-15.
- Bailey, S. (1994). The CCP4 Suite - Programs for Protein Crystallography. *Acta Crystallographica Section D-Biological Crystallography*, 50(5):760-763.
- Becke, A. (1993). Density-Functional ThermochemistryIII. the Role of Exact Exchange. *Journal of Chemical Physics*, 98(7):5648-5652.
- Blöchl, P. (1994). Projector Augmented-Wave Method. *Physical Review B*, 50(24):17953-17979.

Table 6.19 (continued from previous page)

-
- Brown, G., Albers, J., Fisher, L., Schaefer, S., Lin, J., Kaplan, C., Zhao, X., Bisson, B., Fitzpatrick, V., and Dodge, H. (1990). Regression of Coronary-Artery Disease as a Result of Intensive Lipid-Lowering Therapy in Men With High-Levels of Apolipoprotein-B. *New England Journal of Medicine*, 323(19):1289-1298.
- Coussot, P., Leonov, A., and Piau, J. (1993). Rheology of Concentrated Dispersed Systems in a Low-Molecular-Weight Matrix. *Journal of Non-Newtonian Fluid Mechanics*, 46(2-3):179-217.
- Delerue, C., Allan, G., and Lannoo, M. (1993). Theoretical Aspects of the Luminescence of Porous Silicon. *Physical Review B*, 48(15):11024-11036.
- Dixon, R., Brown, S., Houghton, R., Solomon, A., Trexler, M., and Wisniewski, J. (1994). Carbon Pools and Flux of Global Forest Ecosystems. *Science*, 263(5144):185-190.
- Duschek, W., Kleinrahm, R., and Wagner, W. (1990). Measurement and Correlation of the (Pressure, Density, Temperature) Relation of Carbon-Dioxide .1. the Homogeneous Gas and Liquid Regions in the Temperature-Range from 217-K to 340-K At Pressures Up To 9 Mpa. *Journal of Chemical Thermodynamics*, 22(9):827-840.
- Eldeiry, W., Tokino, T., Velculescu, V., Levy, D., Parsons, R., Trent, J., Lin, D., Mercer, W., Kinzler, K., and Vogelstein, B. (1993). WAF1, a Potential Mediator of P53 Tumor Suppression. *Cell*, 75(4):817-825.
- Fang, Z., Liu, Z., and Yao, K. (1994). Theoretical-Model and Numerical-Calculations for a Quasi-One-Dimensional Organic Ferromagnet. *Physical Review B*, 49(6):3916-3919.
- Fischman, D., et al. (1994). A Randomized Comparison of Coronary-Stent Placement and Balloon Angioplasty in the Treatment of Coronary-Artery Disease. *New England Journal of Medicine*, 331(8):496-501.
- Frabetti, P., et al.(1992). Description and Performance of the Fermilab-E687 Spectrometer. *Nuclear Instruments & Methods in Physics Research Section A-Accelerators Spectrometers Detectors and Associated Equipment*, 320(3):519-547.
- Frostig, Y., Baruch, M., Vilnay, O., and Sheinman, I. (1992). High-Order Theory for Sandwich-Beam Behavior With Transversely Flexible Core. *Journal of Engineering Mechanics-Asce*, 118(5):1026-1043.
- Gasser, R., Chilton, N., Hoste, H., and Beveridge, I. (1993). Rapid Sequencing of Rdna from Single Worms and Eggs of Parasitic Helminths. *Nucleic Acids Research*, 21(10):2525- 2526.
- Gavrieli, Y., Sherman, Y., and Bensasson, S. (1992). Identification of Programmed Cell-Death Insitu via Specific Labeling of Nuclear-Dna Fragmentation. *Journal of Cell Biology*, 119(3):493-501.
- Hafner, H., Riecherrossler, A., Hambrecht, M., Maurer, K., Meissner, S., Schmidtke, A., Fatkenheuer, B., Loffer, W., and Vanderheiden, W. (1992). Iraos - an Instrument for the Assessment of Onset and Early Course of Schizophrenia. *Schizophrenia Research*, 6(3):209-223.
- Hodges, J., Patterson, K., Oxbury, S., and Funnell, E. (1992). Semantic Dementia - Progressive Fluent Aphasia With Temporal-Lobe Atrophy. *Brain*, 115(6):1783-1806.
- Huntington, S. (1993). The Clash of Civilizations. *Foreign Affairs*, 72(3):22-49.
- Iijima, S. (1991). Helical Microtubules of Graphitic Carbon. *Nature*, 354(6348):56-58.
- Inoue, A., Yamaguchi, H., Zhang, T., and Masumoto, T. (1990). Al-La-Cu amorphous-alloys with a wide supercooled liquid region. *Materials Transactions Jim*, 31(2):104-109.
-

Table 6.19 (continued from previous page)

-
- Jones, T., Zou, J., Cowan, S., and Kjeldgaard, M. (1991). Improved Methods for Building Protein Models in Electron-Density Maps and the Location of Errors in These Models. *Acta Crystallographica Section A*, 47(2):110-119.
- Keller, G., Kennedy, M., Papayannopoulou, T., and Wiles, M. (1993). Hematopoietic Commitment During Embryonic Stem-Cell Differentiation in Culture. *Molecular and Cellular Biology*, 13(1):473-486.
- Kessler, R., McGonagle, K., Zhao, S., Nelson, C., Hughes, M., Eshleman, S., Wittchen, H., and Kendler, K. (1994). Lifetime and 12-Month Prevalence of DSM-III-R Psychiatric Disorders in the United-States - Results from the National-Comorbidity-Survey. *Archives of General Psychiatry*, 51(1):8-19.
- Kiranoudis, C., Maroulis, Z., Tsami, E., and Marinoskouris, D. (1993). Equilibrium Moisture-Content and Heat of Desorption of Some Vegetables. *Journal of Food Engineering*, 20(1):55-74.
- Koudstaal, P., et al. (1993). Secondary Prevention in Nonrheumatic Atrial-Fibrillation after Transient Ischemic Attack or Minor Stroke. *Lancet*, 342(8882):1255-1262.
- Kraulis, P. (1991). MOLSCRIPT - a program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography*, 24(5):946-950.
- Kubota, Y., et al. (1992). The Cleo-II Detector. *Nuclear Instruments & Methods in Physics Research Section A-Accelerators Spectrometers Detectors and Associated Equipment*, 320(1-2):66-113.
- Kumbharkhane, A., Puranik, S., and Mehrotra, S. (1991). Dielectric-Relaxation of Tert Butyl Alcohol-Water Mixtures Using A Time-Domain Technique. *Journal of the Chemical Society-Faraday Transactions*, 87(10):1569-1573.
- Larsson, A., Stenberg, L., and Lidin, S. (1994). The Superstructure of Domain-Twinned Eta'-Cu₆Sn₅. *Acta Crystallographica Section B-Structural Science*, 50(6):636-643.
- Laskowski, R., MW, M., Moss, D., and Thornton, J. (1993). PROCHECK - a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26(2):283-291.
- Lauck, L., Vasconcellos, A., and Luzzi, R. (1990). A Nonlinear Quantum Transport-Theory. *Physica A-Statistical Mechanics and its Applications*, 168(2):789-819.
- Lebel, C., Ali, S., McKee, M., and Bondy, S. (1990). Organometal-Induced Increases in Oxygen Reactive Species - the Potential of 2',7'-Dichlorofluorescein Diacetate as an Index of Neurotoxic Damage. *Toxicology and Applied Pharmacology*, 104(1):17-24.
- Manabe, T., Wyllie, D., Perkel, D., and Nicoll, R. (1993). Modulation of Synaptic Transmission and Long-Term Potentiation - Effects on Paired-Pulse Facilitation and Epsc Variance in the Ca1 Region of the Hippocampus. *Journal of Neurophysiology*, 70(4):1451-1459.
- Mccormick, D. and Pape, H. (1990). Properties of A Hyperpolarization-Activated Cation Current and its Role in Rhythmic Oscillation in Thalamic Relay Neurons. *Journal of Physiology-London*, 431:291-318.
- Middlemas, D., Lindberg, R., and Hunter, T. (1991). TRKB, a neural receptor proteintyrosine kinase - evidence for a full-length and 2 truncated receptors. *Molecular and Cellular Biology*, 11(1):143-153.
- Miller, A., Baines, C., To, T., and Wall, C. (1992). Canadian-National-Breast-Screening-Study .1. Breast-Cancer Detection and Death Rates Among Women Aged 40 to 49 Years. *Canadian Medical Association Journal*, 147(10):1459-1476.
-

Table 6.19 (continued from previous page)

-
- Moncada, S., Palmer, R., and Higgs, E. (1991). Nitric-oxide - physiology, pathophysiology, and pharmacology. *Pharmacological Reviews*, 43(2):109-142.
- Ohara, P., Sheppard, P., Thogersen, H., Venezia, D., Haldeman, B., Mcgrane, V., Houamed, K., Thomsen, C., Gilbert, T., and Mulvihill, E. (1993). The Ligand-Binding Domain in Metabotropic Glutamate Receptors Is Related to Bacterial Periplasmic Binding-Proteins. *Neuron*, 11(1):41-52.
- O'Regan, B. and Grätzel, M. (1991). A low-cost, high-efficiency solar-cell based on dye-sensitized colloidal TiO₂ films. *Nature*, 353(6346):737-740.
- Pause, A. and Sonenberg, N. (1992). Mutational Analysis of A Dead Box Rna Helicase - the Mammalian Translation Initiation-Factor Eif-4a. *Embo Journal*, 11(7):2643-2654.
- Perdew, J. and Wang, Y. (1992). Accurate and simple analytic representation of the electron-gas correlation-energy. *Physical Review B*, 45(23):13244-13249.
- ProbstinÅeld, J. (1991). Prevention of Stroke By Antihypertensive Drug-Treatment in Older Persons With Isolated Systolic Hypertension - Final Results of the Systolic Hypertension in the Elderly Program (Shep). *Jama-Journal of the American Medical Association*, 265(24):3255-3264.
- Riggs, B., Hodgson, S., Ofallon, W., Chao, E., Wahner, H., Muhs, J., Cedel, S., and Melton, L. (1990). Effect of Fluoride Treatment on the Fracture Rate in Postmenopausal Women With Osteoporosis. *New England Journal of Medicine*, 322(12):802-809.
- Rudman, D., Feller, A., Nagraj, H., Gergans, G., Lalitha, P., Goldberg, A., Schlenker, R., Cohn, L., Rudman, I., and Mattson, D. (1990). Effects of Human Growth-Hormone in Men Over 60 Years Old. *New England Journal of Medicine*, 323(1):1-6.
- Ryder, R. (1993). Natural Family-Planning - Effective Birth-Control Supported By the Catholic-Church. *British Medical Journal*, 307(6906):723-726.
- Sakai, T., Miyamura, H., Kuriyama, N., Kato, A., Oguro, K., Ishikawa, H., and Iwakura, C. (1990). The Influence of Small Amounts of Added Elements on Various Anode Performance-Characteristics for LaNi_{2.5}Co_{2.5}-Based Alloys. *Journal of the Less-Common Metals*, 159(1-2):127-139.
- Shamoon, H., et al. (1993). The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long-Term Complications in Insulin-Dependent Diabetes-Mellitus. *New England Journal of Medicine*, 329(14):977-986.
- Sheldrick, G. (1990). Phase annealing in SHELX-90 - direct methods for larger structures. *Acta Crystallographica Section A*, 46(6):467-473.
- Stampfer, M., Colditz, G., Willett, W., Manson, J., Rosner, B., Speizer, F., and Hennekens, C. (1991). Postmenopausal Estrogen Therapy and Cardiovascular-Disease - 10-Year Follow-Up from the Nurses Health Study. *New England Journal of Medicine*, 325(11):756-762.
- Stave, M., Sanders, D., Raeker, T., and Depristo, A. (1990). Corrected Effective Medium Method .5. Simplifications for Molecular-Dynamics and Monte-Carlo Simulations. *Journal of Chemical Physics*, 93(6):4413-4426.
- Svoboda, J., Riedel, H., and Zipse, H. (1994). Equilibrium Pore Surfaces, Sintering Stresses and Constitutive-Equations for the Intermediate and Late Stages of Sintering .1. Computation of Equilibrium Surfaces. *Acta Metallurgica Et Materialia*, 42(2):435-443.
-

Table 6.19 (continued from previous page)

Swedberg, K., Held, P., Kjekshus, J., Rasmussen, K., Ryden, L., and Wedel, H. (1992). Effects of the Early Administration of Enalapril on Mortality in Patients With Acute Myocardial-Infarction - Results of the Cooperative New Scandinavian Enalapril Survival Study-II (Consensus-II). *New England Journal of Medicine*, 327(10):678-684.

Thompson, J., Higgins, D., and Gibson, T. (1994). Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research*, 22(22):4673-4680.

Ware, J. and Sherbourne, C. (1992). The MOS 36-item short-form health survey (SF-36) .1. conceptual-framework and item selection. *Medical Care*, 30(6):473-483.

Whitesides, G. and Laibinis, P. (1990). Wet Chemical Approaches to the Characterization of Organic-Surfaces - Self-Assembled Monolayers, Wetting, and the Physical Organic-Chemistry of the Solid Liquid Interface. *Langmuir*, 6(1):87-96.

Williams, J., Kubelik, A., Livak, K., Rafalski, J., and Tingey, S. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic-markers. *Nucleic Acids Research*, 18(22):6531-6535.

Zhang, Y., Proenca, R., Maffei, M., Barone, M., Leopold, L., and Friedman, J. (1994). Positional Cloning of the Mouse Obese Gene and its Human Homolog. *Nature*, 372(6505):425-432.

6.A.6 Key figures for the papers in the test set

In this section the outcomes of applying the algorithms to the test set of 60 papers are further broken down. Table 6.20 shows some bibliographic key figures for the breakout papers in the test set. Some further characteristics of the algorithms are presented in Table 6.21, where we differentiate for each algorithm between the papers that are selected (breakout papers) and those that are not selected (non-breakout papers). The table shows the percentages of papers belonging to the Top 3%, based on the number of citations a paper received. Table 6.22 shows for the papers in the test set the number of times a paper is cited by a review paper, and provides also some commenting information.

Table 6.20: Key figures for the various algorithms applied to the test set with 60 papers

Algorithm	Number of breakout papers	Range of times-cited in 24 months by articles and letters	Occuring in Nature's Top-100 most cited list	Cited by patent families (PATSTAT spring 2014 edition)	Cited in Social Media (2012-2014)	Range of number of times cited by review papers
ARI	15	17-51	0	7	0	18-150
CDI	33	24-579	11	25	10	48-2133
RII	26	65-579	8	20	9	48-2133
DII	16	3-19	0	2	0	1-38
RNI	12	30-199	0	6	4	1-684

6.A Supplementary material

Table 6.21: Characteristics of the algorithms by differentiating the papers from the test set in breakout papers and non-breakout papers. Shown are the percentages of papers in each of the two groups that belong to the Top 3% percentile of distribution of the number of citations received from the different sources by all papers (letters and articles) published in 1990-1994 and covered in the WoS database.

Algorithm	Cited by review papers		Cited by patents		Times cited within 24 months	
	Breakout paper	Non-breakout paper	Breakout paper	Non-breakout paper	Breakout paper	Non-breakout paper
<i>(Percentage of papers in Top 3%)</i>						
ARI	93%	64%	7%	27%	100%	67%
CDI	100%	39%	41%	0%	100%	46%
RHI	100%	50%	42%	6%	100%	56%
DHI	6%	95%	0%	30%	6%	100%
RNI	92%	67%	17%	23%	100%	69%

Table 6.22: Number of times a paper in the test set is cited by a review paper

Paper	Times cited by review papers until 1st Quarter of 2015	Notes
Adeva et al. (1990)	38	Top 1.5%
Ahmet (1991)	37	Top 1.5%
Altschul et al. (1990)	1,155	1990 paper most cited by review papers
Angell (1991)	72	Top 0.5%
"Anonymous" (1992)	34	Top 2 %
"Anonymous" (1991)	350	Top 0.5%
Bailey (1994)	149	Top 0.5%
Becke (1993)	1,664	1993 paper that belongs to the most cited by review papers within the top 0.5%
Blöchl et al. (1994)	162	Top 0.5%
Brown et al. (1990)	385	Top 0.5%
Coussot et al. (1993)	4	Top 40%
Delerue et al. (1993)	37	Top 2 %
Dixon et al. (1994)	98	Top 0.5%

Table 6.22 (continued from previous page)

Paper	Times cited by review papers until 1st Quarter of 2015	Notes
Duschek et al. (1990)	9	Top 20%
Eldeiry et al. (1993)	942	1993 paper that belongs to the most cited by review papers within the top 0.5%
Fang et al. (1994)	1	Low cited
Fischman et al. (1994)	378	Top 0.5%
Frabetti et al. (1992)	7	Top 20%
Frostig et al. (1992)	2	Low cited
Gasser et al. (1993)	11	Top 20%
Gavrieli et al. (1992)	457	Top 0.5%
Hafner et al. (1992)	38	Top 2 %
Hodges et al. (1992)	150	Top 0.5%
Huntington et al. (1993)	97	Top 0.5%
Iijima (1991)	1,509	1991 paper most cited by review papers
Inoue et al. (1990)	6	Top 30%
Jones et al. (1991)	180	Top 0.5%
Keller et al. (1993)	100	Top 0.5%
Kessler et al. (1994)	871	1994 paper that belongs to the most cited by review papers within the top 0.5%
Kiranoudis et al. (1993)	3	Top 50%
Koudstaal et al. (1993)	300	Top 0.5%
Kraulis (1991)	832	1991 paper that belongs to the most cited by review papers within the top 0.5%
Kubota et al. (1992)	18	Top 10%
Kumbharkhane et al. (1991)		Not cited
Larsson et al. (1994)	5	Top 30%
Lauck et al. (1990)	16	Top 10%
Lebel et al. (1990)	15	Top 10%

6.A Supplementary material

Table 6.22 (continued from previous page)

Paper	Times cited by review papers until 1st Quarter of 2015	Notes
Manabe et al. (1993)	35	Top 20%
Mccormick and Pape (1990)	118	Top 0.5%
Middlemas et al. (1991)	91	Top 0.5%
Miller et al. (1992)	48	Top 1%
Ohara et al. (1993)	736	1991 paper that belongs to the most cited by review papers within the top 0.5%
O'Regan and Grätzel (1991)	141	Top 0.5%
Pause and Sonenberg (1992)	58	Top 1%
Probstfield (1991)	684	1991 paper that belongs to the most cited by review papers within the top 0.5%
Riggs et al. (1990)	188	Top 0.5%
Rudman et al. (1990)	282	Top 0.5%
Ryder et al. (1993)	1	Low cited
Sakai et al. (1990)	4	Top 40%
Shamoon et al. (1993)	2,133	1993 paper most cited by review papers
Sheldrick (1990)	178	Top 0.5%
Stampfer et al. (1991)	275	Top 0.5%
Stave et al. (1990)	6	Top 30%
Svoboda et al. (1994)	7	Top 20%
Swedberg et al. (1992)	197	Top 0.5%
Thompson et al. (1994)	1,154	1994 paper that belongs to the most cited by review papers within the top 0.5%
Ware and Sherbourne (1992)	1,102	1992 paper most cited by review papers
Whitesides and Laibinis (1990)	80	Top 0.5%
Williams et al. (1990)	379	Top 0.5%
Zhang et al. (1994)	1,607	1994 paper most cited by review papers

Table 6.23: Classification of the results based on the total number of citations received in comparison with the distribution of the citation scores for all papers ('letters' and 'articles') published in the period 1990–1994 and covered in the WoS database; percentile boundaries are presented between parentheses.

Algorithm	Papers recognized	Not cited	Low (<70%)	Medium (70–97%)	High (97–99%)	Very High (≥ 99%)
Citations from patents						
ARI	15	18	13	13	11	10
CDI	32	8	2	9	5	8
RII	26	6	2	7	4	7
DII	16	14	1	1	0	0
RNI	12	6	1	3	1	1
Citations from review papers						
ARI	15	0	0	1	5	9
CDI	32	0	0	0	1	31
RII	26	0	0	0	1	25
DII	16	1	5	9	1	0
RNI	12	0	1	0	1	10
Citations from original research papers within 24 months						
ARI	15	0	0	0	2	13
CDI	32	0	0	0	1	31
RII	26	0	0	0	0	26
DII	16	0	2	13	1	0
RNI	12	0	0	0	0	12

The scores for each of the papers in the test set on the dimensions (1) number of citations from patents, (2) number of citations from review papers, and (3) the number of citations received within 24 months after publication is presented in 6.23. As the distribution of the number of citations is very skewed, we use the following classification scheme: 'Not cited', 'Low' (<70%), 'Medium' (70–97%), 'High' (97–99%), and 'Very High' (≥ 99%). The numbers between parentheses are the percentile boundaries.

6.A.7 Breakout detection specificity of the algorithms

The recall rate for the algorithms is presented for various threshold values. The specificity or recall rates of the algorithms when applied to the two datasets as a function of the imposed thresholds are presented in Table 6.24 and Table 6.25. Table 6.24 focuses on the Categories dataset and Table 6.25 on the Clusters dataset. In figures Figure 6.10 and Figure 6.11 this information is presented in graphical form.

Table 6.24: Categories: breakout detection specificity (recall rate) of the algorithms. Ratios are based on the absolute numbers in Table 6.13

	ARI	CDI _{SC}	RII	DII	RNI	REDNER's algorithm
Threshold (Number of times cited within 24 months)						
≥ 1	0.10%	0.50%	1.40%	0.23%	0.01%	2.43%
≥ 2	0.10%	0.51%	1.40%	0.23%	0.01%	2.44%
≥ 4	0.11%	0.54%	1.49%	0.17%	0.01%	2.58%
≥ 8	0.17%	0.81%	2.26%	0.05%	0.01%	3.75%
≥ 16	0.38%	2.41%	6.68%	0.00%	0.04%	9.46%
≥ 32	0.27%	9.17%	26.08%	0.00%	0.10%	27.59%
≥ 64	0.00%	30.84%	96.50%	0.00%	0.43%	63.92%
≥ 128	0.00%	69.57%	100.00%	0.00%	0.74%	94.81%
≥ 256	0.00%	96.36%	100.00%	0.00%	0.00%	100.00%
≥ 512	0.00%	100.00%	100.00%	0.00%	0.00%	100.00%
≥ 1024						

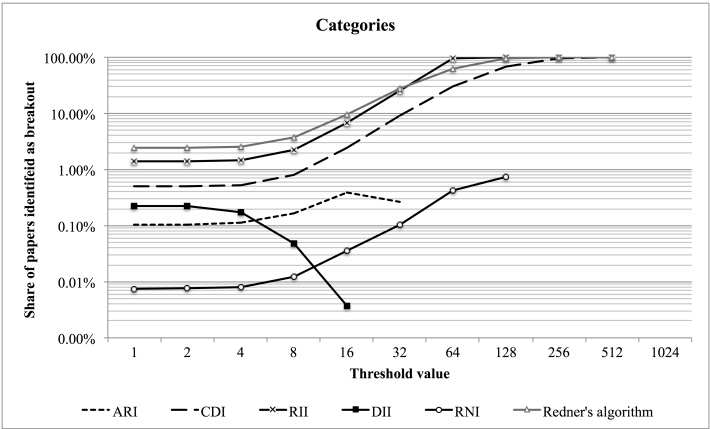


Figure 6.10: Categories: share of documents identified as breakout as a function of the threshold value

Table 6.25: Clusters: breakout detection specificity (recall rate) of the algorithms. Ratios are based on the absolute numbers in Table 6.14

	ARI	CDI _{dc}	RII	DII	RNI	REDNER's algorithm
Threshold (Number of times cited within 24 months)						
≥ 1	0.03%	6.27%	1.63%	0.31%	0.00%	2.94%
≥ 2	0.03%	6.29%	1.64%	0.31%	0.00%	2.95%
≥ 4	0.03%	6.69%	1.74%	0.24%	0.00%	3.13%
≥ 8	0.04%	9.75%	2.54%	0.05%	0.01%	4.23%
≥ 16	0.12%	23.41%	6.78%	0.00%	0.02%	10.02%
≥ 32	0.42%	51.84%	26.19%	0.00%	0.05%	28.04%
≥ 64	0.00%	79.29%	97.64%	0.00%	0.24%	65.07%
≥ 128	0.00%	91.01%	100.00%	0.00%	0.75%	95.13%
≥ 256	0.00%	100.00%	98.15%	0.00%	0.00%	100.00%
≥ 512	0.00%	100.00%	100.00%	0.00%	0.00%	100.00%
≥ 1024						

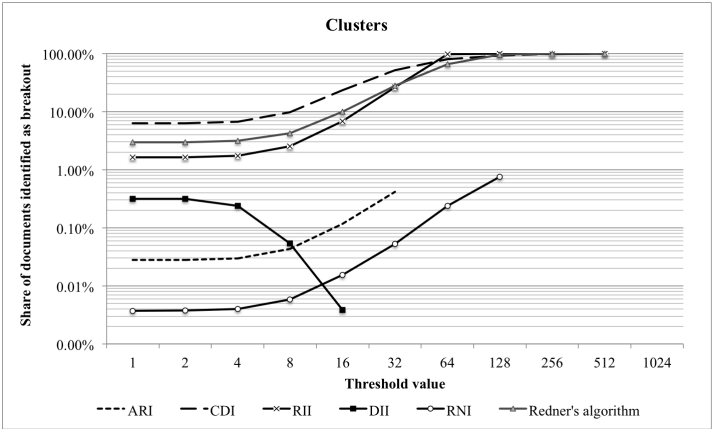


Figure 6.11: Clusters: share of documents identified as breakout as a function of the threshold value

Table 6.26: Detection probabilities for each algorithm on the two datasets

	Categories	Clusters
ARI	1.041×10^{-3}	2.793×10^{-4}
CDI	5.032×10^{-3}	6.273×10^{-2}
RII	1.397×10^{-2}	1.630×10^{-2}
DII	2.276×10^{-3}	3.137×10^{-3}
RNI	7.494×10^{-5}	3.724×10^{-5}
REDNER	2.425×10^{-2}	2.938×10^{-2}

Table 6.27: Categories: expected breakout recall

	ARI	CDI _{sc}	RII	DII	RNI	REDNER's algorithm
Threshold (Number of times cited within 24 months)						
≥ 1	264	1,276	3,543	577	19	6,150
≥ 2	262	1,269	3,525	574	18	6,119
≥ 4	247	1,197	3,325	541	17	5,772
≥ 8	163	788	2,190	356	11	3,802
≥ 16	55	266	740	120	3	1,285
≥ 32	14	68	189	30	1	329
≥ 64	3	15	41	6	0	72
≥ 128	0	2	7	1	0	13
≥ 256	0	0	0	0	0	1
≥ 512	0	0	0	0	0	0
≥ 1024	0	0	0	0	0	0

6.A.8 Detection probability and expected breakout recall rate of the algorithms

In this section the performance of the algorithms, including Redner's algorithm, on the datasets Categories and Clusters is compared with what might be expected. The 'breakout detection probability' of an algorithm is the probability that the algorithm selects a paper as breakout paper. For both Categories and Clusters, these probabilities are calculated and shown in Table 6.26. We calculated the 'expected breakout recall' of the algorithms by multiplying the appropriate detection probability with the number of papers in a dataset after applying a particular threshold. The expected breakout recall of the algorithms for the two datasets as function of the threshold is presented in Tables 6.27 and 6.29. The difference between the identified number and the expected number of breakout papers as a percentage of the expected number is presented in Tables 6.28 and 6.30, and is also shown in Figure 6.12 and Figure 6.13.

Table 6.28: Categories: difference between the identified number and the expected number of breakout papers as percentage of the expected number of breakout papers.

	ARI	CDI _{dc}	RII	DII	RNI	REDNER's algorithm
Threshold (Number of times cited within 24 months)						
≥ 1	0%	0%	0%	0%	0%	0%
≥ 2	1%	1%	1%	0%	6%	1%
≥ 4	7%	7%	7%	-24%	12%	6%
≥ 8	62%	62%	62%	-79%	73%	55%
≥ 16	271%	379%	379%	-98%	533%	290%
≥ 32	157%	1,732%	1,775%	-100%	1,300%	1,039%
≥ 64	-100%	6,060%	6,951%	-100%		2,560%
≥ 128		18,650%	7,600%	-100%		3,831%
≥ 256						5,400%
≥ 512						
≥ 1024						

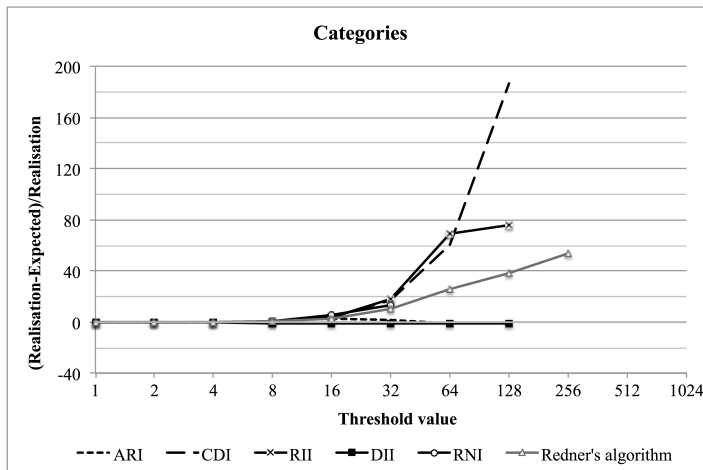


Figure 6.12: Categories: difference between the identified number and the expected number of breakout papers as percentage of the expected number of breakout papers

Table 6.29: Clusters: expected breakout recall

	ARI	CDI _{sc}	RII	DII	RNI	REDNER'S algorithm
Threshold (Number of times cited within 24 months)						
≥ 1	60	13,477	3,501	674	8	6,311
≥ 2	59	13,410	3,483	670	7	6,280
≥ 4	56	12,650	3,286	632	7	5,923
≥ 8	37	8,332	2,164	416	4	3,901
≥ 16	12	2,817	731	140	1	1,319
≥ 32	3	721	187	36	0	338
≥ 64	0	159	41	7	0	74
≥ 128	0	28	7	1	0	13
≥ 256	0	2	0	0	0	1
≥ 512	0	0	0	0	0	0
≥ 1024	0	0	0	0	0	0

Table 6.30: Clusters: difference between the identified number and the expected number of breakout papers as percentage of the expected number of breakout papers.

	ARI	CDI _{dc}	RII	DII	RNI	REDNER'S algorithm
Threshold (Number of times cited within 24 months)						
≥ 1	0%	0%	0%	0%	0%	0%
≥ 2	2%	0%	1%	0%	14%	0%
≥ 4	7%	7%	7%	-24%	14%	6%
≥ 8	62%	61%	62%	-82%	100%	50%
≥ 16	400%	329%	379%	-99%	700%	293%
≥ 32	1,767%	861%	1,772%	-100%		1,009%
≥ 64		1,359%	6,868%	-100%		2,473%
≥ 128		1,636%	7,529%	-100%		3,808%
≥ 256		2,600%				5,300%
≥ 512						
≥ 1024						

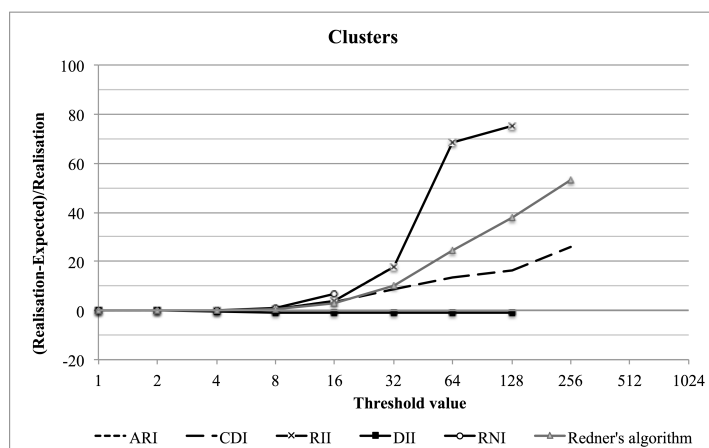


Figure 6.13: Categories: difference between the identified number and the expected number of breakout papers as percentage of the expected number of breakout papers

Table 6.31: Labels of distribution classes used in tables 6.32–6.40

Label	Criterion	Description
Top 1%	[99.0-100.0)%	Top 1 %
Top 5%	[95.0-99.0)%	In Top 5% but not in Top 1%
Top 10%	[90.0-95.0)%	In Top 10% but not in Top 5%
< Top 10%	[0.0-90.0)%	Outside Top 10%

6.A.9 Breakout classification

In this section the ‘breakout character’ of the results after applying the algorithms is presented. We use the following classification scheme (Table 6.31) for the number of citations received by a paper. On the basis of this classification scheme we obtain Tables 6.32–6.40 in which the number of citations from review papers is the first dimension (rows) and the number of citation received from patents the second dimension (columns). In Table 6.35 we show the effect of the algorithms by comparing the results presented in Table 6.34 with identified breakouts with the distribution for all papers (Table 6.33). A ‘+’ in a cell means that after applying the algorithms the concentration of the papers in this cell is higher than in comparable cell in the total dataset (Table 6.33), and a ‘-’ means that the concentration is lower. The algorithms succeed in selecting papers after 20+ years belong to the papers highly cited by review papers, and by patents.

6.A Supplementary material

Table 6.32: Categorisation of all 1990–1994 papers (articles, letters) cited within 24 months, along the two dimensions ‘cited by review papers’ and ‘cited by patents’. Based on citation counts from the moment of publication until 2015.

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	14,262	2,402	1,679	508
Top 5%	146,335	7,416	3,326	544
Top 10%	524,749	9,472	3,214	420
< Top 10%	1,073,108	13,137	4,102	571

Table 6.33: Distribution of all 1990–1994 papers (articles, letters) cited within 24 months, along the two dimensions ‘cited by review papers’ and ‘cited by patents’. Based on citation counts from the moment of publication until 2015.

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	0.79%	0.13%	0.09%	0.03%
Top 5%	8.11%	0.41%	0.18%	0.03%
Top 10%	29.07%	0.52%	0.18%	0.02%
< Top 10%	59.44%	0.73%	0.23%	0.03%

Table 6.34: Categorisation of all 1990–1994 papers (articles, letters) including papers not cited^a within 24 months, along the two dimensions ‘cited by review papers’ and ‘cited by patents’. Based on citation counts from the moment of publication until 2015.

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	14,263	2,403	1,680	511
Top 5%	146,378	7,428	3,327	548
Top 10%	524,874	9,488	3,223	420
< Top 10%	2,690,150	13,163	4,110	572

^a These papers are included in the cell (<Top 10%, <Top 10%)

Table 6.35: Distribution of all 1990-1994 papers (articles, letters) including papers not cited^b within 24 months, along the two dimensions 'cited by review papers' and 'cited by patents'. Based on citation counts from the moment of publication until 2015.

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	0.42%	0.07%	0.05%	0.01%
Top 5%	4.28%	0.22%	0.10%	0.02%
Top 10%	15.34%	0.28%	0.09%	0.01%
< Top 10%	78.60%	0.38%	0.12%	0.02%

^b These papers are included in the cell (<Top 10%, <Top 10%)

Table 6.36: Categorisation of all 1990-1994 papers (articles, letters) identified as a breakout paper, along the two dimensions 'cited by review papers' and 'cited by patents'. Based on citation counts from the moment of publication until 2015.

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	1,947	1162	916	267
Top 5%	1,744	745	379	70
Top 10%	1,104	297	128	17
< Top 10%	1,831	376	109	14

Table 6.37: Distribution of all 1990-1994 papers (articles, letters) identified as a breakout paper, along the two dimensions 'cited by review papers' and 'cited by patents'. Based on citation counts from the moment of publication until 2015.

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	17.53%	10.46%	8.25%	2.40%
Top 5%	15.70%	6.71%	3.41%	0.63%
Top 10%	9.94%	2.67%	1.15%	0.15%
< Top 10%	16.49%	3.39%	0.98%	0.13%

Table 6.38: Effect of applying the algorithms to the complete dataset with papers (articles, letters) from 1990–1994

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	+	+	+	+
Top 5%	-	+	+	+
Top 10%	-	+	+	+
< Top 10%	-	+	+	+

Table 6.39: Categorisation along the two dimensions of all 1990–1994 papers (articles, letters) that are not cited within 24 months, but are cited at least once at a later moment. Based on citation counts from the moment of publication until 2015.

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	34	1	2	
Top 5%	2,788	63	27	4
Top 10%	53,586	509	141	
< Top 10%	346,343	3,661	1,029	141

Table 6.40: Distribution along the two dimensions of all 1990–1994 papers (articles, letters) that are not cited within 24 months, but are cited at least once at a later moment. Based on citation counts from the moment of publication until 2015.

Citations by review papers	Citations by patents			
	< Top 10%	Top 10%	Top 5%	Top 1%
Top 1%	0.01%	0.00%	0.00%	0.00%
Top 5%	0.68%	0.02%	0.01%	0.00%
Top 10%	13.12%	0.12%	0.03%	0.00%
< Top 10%	84.83%	0.90%	0.25%	0.03%

Conclusions, a definition of ‘breakthrough’, future prospects

7.1 General concluding remarks

This study set out to tackle the primary research question: “*Is it possible to design, develop, implement, and test an analytical framework and measurement model for early detection of breakthroughs in worldwide science?*” Facing this methodological challenge, first of all requires a definition of the key concept: ‘scientific breakthrough’. To develop such a framework and model, Hollingsworth’s definition (Hollingsworth, 2008, p.317) “...*A major breakthrough or discovery is a finding or process, often preceded by numerous small advances, which leads to a new way of thinking about a problem ...*” was used as a general guide to further operationalize the concept and to select historical cases for further research and development.

After having concluded a series of case studies, and tested the detection algorithms on real-life data, the answer to this prime question that emerges from this PhD study is conditionally affirmative – in other words a ‘yes, but’.

Regarding the ‘yes’ part of the answer, in Section 1.3.1 the question is addressed whether or not it would be possible to identify a *breakthrough* with only bibliographic information at one’s disposal. Scholars like Redner (2005), Ponomarev et al. (2014b), and Schneider and Costas (2015) also focus on citation impact profiles and consider a highly cited publication a possible signal of a breakthrough discovery. However, they fail to take the dynamic behaviour of science, and more in particular the unpredictable influence of a discovery on the science community, into account. Moreover, by focussing on individual publications, those major discoveries that are presented in several closely related papers published in a relative short time period are being missed. In this study the dynamic behaviour is used as the computational basis for the algorithms.

By focussing on characteristic patterns within citation profiles of prior-known breakthrough discoveries the problem of a missing generally accepted definition of those breakthroughs is partially circumvented. The citation patterns reflect the impact of a discovery on its scientific environment, and provide verifiable empirical information that can be used to design, construct and 'train' computerised algorithms for identifying scholarly publications at an early stage. These high-profile 'exemplars' of breakthroughs are then used as 'analytical templates' to search and find related cases, hitherto undetected 'breakout' publications, in the worldwide research literature.

The five 'early detection' algorithms succeeded in finding thousands of breakouts. Each algorithm captures a different breakthrough characteristic. Although the underlying search queries and selection parameters were set in advance, and in accordance with expert analysis of case study findings, the application of the detection algorithms within the final validation study are disconnected from human decision making. Such an unbiased and transparent selection generates outcomes that are reproducible and independent from any specific information on developments in science - other than the prior-imputed general 'template' of breakout.

Expert opinion and sufficiently long time-spans are needed to ascertain if a breakout is in fact a breakthrough or not. The validation study, on a small sample of the breakout publications in 1990-1994, proved beyond a doubt that the algorithms are able to detect breakthroughs. A very significant fraction of the sampled breakouts were either Nobel Prize winning discoveries, highly-cited in review papers, cited in patents, or (still) cited within social media. Reflecting on the findings of the case studies and concluding validation study, the Hollingsworth's concept of 'scientific breakthrough' can be further empirically operationalized:

A major breakthrough or discovery is an acknowledged discovery, or a cluster of closely related discoveries, published in the open literature within a relative short period of time, of which the citation impact on other researchers marks an identifiable discontinuity in the cognitive development of a science system. This discontinuity manifests itself as a sudden and significant increase of the publication activity directly related to the discovery or discoveries, and a time-delayed measurable impact on relevant user communities.

The findings so far still leave several important methodological issues unresolved, which brings us to the 'but' part of the general conclusions. First, the detection algorithms are unable to differentiate between *hypes* and real breakthroughs given the short time-period taken into account (i.e., 2 or 3 years after publication of the discovery). Secondly, the study's design and its outcomes invite the almost inevitable question: 'is there a *best* algorithm in terms of its hit rate?' The answer, as far as this can be concluded from this study is 'no', as each of the algorithms focuses on different characteristic patterns in the citation profiles. The 'best' algorithm depends on the type of breakout or breakthrough one is looking for. Thirdly, the set of validation criteria is neither

necessarily exhaustive nor sufficiently precise for detecting breakthroughs. Notably we use granted Noble prizes as an ultimate indicator. Nobel Prize decisions are the outcome of a complex process with no guarantees that the world's 'top class' researches will always get awarded. Moreover, laureates are also awarded to scientists for interconnected 'oeuvres' of publications, an information source that is not handled by the current detection algorithms (having been designed to search for single publications). Which brings us to the forth cautionary remark: these detection algorithms are one-off 'mechanistic' and deterministic devices for filtering individual research publications. Their success rate might be enhanced through manual, step-wise 'trial and error' optimization of filter parameter settings. As such they have the potential to become user-interactive 'learning algorithms'. Further development could, ultimately, lead to sophisticated artificial intelligence tools of the truly automated 'deep learning' type producing results that can not, easily, be deduced from the data: the computer, somehow, determines what a breakout is.

7.2 Discussion and future prospects

7.2.1 Methodological issues

This empirical study, suffering from inevitable constraints in terms of time and available resources, left several open questions and unresolved problems that were not (sufficiently) addressed and therefore open for further discussion and follow-up work. This section reflects on those topics that are grouped under the general headings of overarching methodological issues.

Methodology related issues

Breakouts, breakthroughs, breakthrough-class? This study tries to come around the issue that a general accepted definition for 'breakthrough' does not exist at this moment by focussing on the response from the scientific community to a discovery. As argued in Section 1.2.3 discoveries vary in impact on science. The scientific community is able to value on the basis of detailed knowledge of a particular science field and usually after a considerable amount of time has passed if a discovery is to be considered a breakthrough. The algorithms developed in this study use short time windows for the analysis of citation patterns and are, in general, unable to distinguish scientific discoveries that stand out from hypes, hoaxes or fraud. By focussing on detection at early-stage the available information to evaluate the impact of a discovery is limited. Publications selected by the algorithms are labelled 'breakouts' to indicate that additional effort is needed to conclude if the discovery is indeed is to be considered a 'breakthrough'. Breakout publications that stand out in the number of citations received from review papers, from patents or both are given an appropriate 'breakthrough by proxy' label that

are defined on page 116. These ‘breakthrough by proxy’ papers form the set of ‘breakthrough-class’ discoveries.

What is the relation of this study with research on breakthrough innovations and emerging technologies? From the start this study focused on discoveries in science to find those that are expected to have an above average impact on science and possibly result in new technologies. Factors that make a scientific discovery to become the start of a new technology are not addressed in this study. Neither does this study address the research area of detecting breakthrough innovations and monitoring emerging technologies.

Is the number of citations at early-stage a measure for the total number of citations a publication will receive? Adams (2005) concludes that there is a correlation between the number of citations a publication receives at early-stage and the number of citations it will receive in the long-term (t_{∞}). In this study (Chapter 6) a similar relation is found for citations a publication receives from review papers. Wang et al. constructed and tested a theoretical model on this relationship, and conclude that the method could be improved by “mechanistic understanding of the factors that govern the research community’s response to a discovery” (Wang et al., 2013, p.132). As is indicated by Wang et al. and also by Bettencourt et al. (2009, p.220) the search should be on the mechanisms underlying discoveries in science.

Are Kuhn and Koshland the only option to classify discoveries? In this study the choice was made to use the typology proposed by Koshland (2007) to classify discoveries. Koshland’s ideas can be linked with Kuhn’s dichotomy of science and in this respect a closed cognitive system is available. Discoveries can be classified on the basis of criteria other than those proposed by Kuhn and Koshland but in this study it was chosen not to do so to avoid additional complexity.

Generalization and source-dependence

Has ‘highly citedness’ been abandoned in this analytical framework? The algorithms only select publications that received at least one citation during the first 24 months after publishing, but contain no other explicit lower citation threshold. To validate the algorithms, two partly overlapping samples of scholarly publications of types ‘article’ and ‘letter’ from the period 1990–1994 were used. The selected publications belong to the top 10% publications that are the highest cited within 24 months⁹⁹ after publication for each year, and each WoS subject category or CWTs document cluster. The top 10% criterion used is field and publication year specific. The rationale behind this criterion is that a discovery can only become a breakthrough if it is recognised by the

⁹⁹47% of all publications of types ‘article’ and ‘letter’ from 1990–1994 are not cited in the first 24 months after publication, and are therefore not included in the selection.

scientific community and thus cited, and probably already within 24 months after publication. Given the skewness of the citation distribution the top 10% percentile seems an appropriate threshold to avoid 'noise'. The top 10% criterion is an often-used in bibliometrics.

How will the detection algorithms perform on general databases with bibliographic data of scholarly publications other than the WoS? The algorithms are designed for the early stage identification of discoveries in science represented by research publications – as indexed by the Thomson Reuters-owned¹⁰⁰ *Web of Science* database (WoS). This WoS specificity raises the question if the detection algorithms can be applied to other databases containing bibliographic data of scholarly publications. The answer is yes if (1) citation relations interlinking the publications are provided, and (2) time stamps that enable systematic tracking and monitoring of temporal developments are available. The aim of the WoS as well as the Elsevier-owned SCOPUS bibliographic database is to provide a representative picture of scientific research over time in all fields of science, therefore the use of either of these databases should give similar results; there is no need to alter the algorithms as both databases provide similar information.

Do retracted publications affect the algorithms? According to Cokol et al. (2008) the retraction of scientific publications is increasing, and the number of retracted papers in MEDLINE^{®101} reached the 1% level in 2006. Steen et al. (2013, Table 1) show that the mean time to retract a publication depends on the reason to retract and ranges from 26 to almost 47 months. Retracted publications do not vanish from the scientific knowledge base, and are still cited even after their retraction (van Noorden, 2011); in only 8% of the citations is the retraction mentioned. Retracted articles live on in personal libraries and on the Internet (Davis, 2012). Retracted publications are therefore in general present as a referenced publication or as a citing document in the first 24–36 month after publication period that is used in this thesis. After the identification of breakout papers a check for retractions should be carried out to prevent such papers to be seen as a breakout.

Is the validation study biased as a result of the used data? In order to validate the algorithms, publications of the type *article* and *letter* published in the period 1990–1994 are used. This period was chosen to allow enough time to analyse a publication's impact on science. It is assumed that the algorithms are time invariant; this is however not further investigated in this study. This point is partly addressed in two of the follow-up studies that use publications

¹⁰⁰The division of Thomson Reuters responsible for the WoS has recently (July 2016) been sold to two investment firms: *Onex Corporation* and *Baring Private Equity Asia*.

¹⁰¹MEDLINE[®] is the U.S. National Library of Medicine (NLM) premier bibliographic database that contains more than 22 million references to journal articles in life sciences with a concentration on biomedicine.

from the period 2007–2011. Publications of types ‘article’ and ‘letter’ are seen as the messengers of original scientific research. It is known, however, that the assignment of document types to individual scholarly publications is not perfect, but the numbers of errors made in this process are considered to be small enough to be neglected.

Does the use of a time window of 24–36 months cause some breakouts to be inadvertently not recognised? In one of the follow-up studies (see the section ‘At what moment in a paper’s life does the breakout character become apparent for the first time?’ on page 176 section ‘ it is found that for almost 92% of the publications that show ‘breakout character’ during the first 10 years after publication this behaviour manifests itself in the first year; therefore focussing on the time period of 24–36 months with the publication date as point of reference seems to be appropriate. Applying the algorithms to also use the situation of a paper after the first year would increase the hit rate with an extra 6.4%.

Using scholarly literature cited in patents Scholarly publications referenced in patent publications occur as so-called non-patent literature references (NPLRs) in the patent publications. The NPLR information is stored in free-format, not-well-structured text strings. These text strings can contain references to all information that, from a patent procedural point of view, is not considered a reference to a patent publication. The references to scholarly publications therefore form a subset of approximately 50% of the NPLRs; parsing and analysing these text strings is in general a complicated, error-prone and time-consuming task. This issue did not pose problems in the case studies as the data set were relatively small and could be inspected manually.

Applicability

Can the algorithms be applied in a meaningful way to networks of other types of information? The algorithms are designed to analyse citation networks of scholarly publications in order to use the algorithms for other types of information the first question that needs to be answered is “Are there in this *new* domain theoretical concepts that can considered to be equivalents of the concepts of Kuhn (1962), Hollingsworth (2008), and Koshland (2007) that form the foundation of the current algorithms?” Furthermore the validity the algorithms and their concepts in the new domain need to be checked. Even if it is concluded that the algorithms may be applied in the new domain they need to be calibrated, and examples of what is considered a breakthrough in this new domain needs to be found for this calibration.

Can the algorithms be applied to information networks consisting of a combination of different types of information? When combining information from multiple data sources containing different types of information

several issues need to be investigated and addressed. For instance combining databases with scholarly information and databases that contain information other types of information, such as patents or so-called altmetrics¹⁰² data, the algorithms definitely need to be changed. Such changes are needed as different domains in general differ in the precise meaning of timestamps – e.g. publication dates –, in links between individual items, and in the way variations of the same name are handled to name a few complicating issues. The name variant issue prevents some potential algorithms from being used in a generally applicable fashion without much extra effort. For instance algorithms using names of affiliations are therefore not implemented in this study, although their potential was identified in Chapter 3.

Do the algorithms have a bias towards technology? The algorithms are designed on the basis of outcomes of the analysis of breakthrough discoveries that led to new technologies. This raises the question “Are the algorithms general applicable or are they restricted to typical technology-related science?” After the research for this study was concluded, the algorithms were applied to all publications of type article and letter from the period 2007–2011 covered in the WoS database. Outcomes of that study show that even for publications in the fields of ‘Arts & Humanities’ and ‘Social Sciences’ the algorithms identified publications as a breakout. The numbers of publications seen as a potential breakthrough are for these ‘non-sciences’ fields low compared to the numbers found for the sciences fields. Probable causes for this observation can be a less relevant or different role of the ‘discovery’ concept in these fields, and also differences in publication and citation behaviour. A ‘sciences-bias’ might implicitly been built into the algorithms as all case studies used are in the fields of Physics, Biochemistry and Life Sciences; research in the Arts & Humanities and the Social Sciences is less tied to technology.

What about the less formal exchange of information? The focus in this study is on formal citation relations. The exchange of information in less formal ways is not taken into account, as this information is not stored in bibliographic databases. Especially within research teams this informal exchange of information might be important but is restrained to the research team. The publication of a research paper can be seen as a moment of synchronisation of (a part of) the knowledge of the research team with the outside scientific world. Altmetrics data sources contain information that is to be considered an additional way of making information known to the outside world. The value of these data for early stage ‘breakthrough’ detection has yet to be ascertained.

¹⁰²“... Briefly defined, alternative metrics, or altmetrics, are indicators of impact and engagement with research that extend beyond traditional citations. Altmetrics measure attention to research in non-traditional sources such as news, blogs, policy documents and social media ...” (<http://www.sheffield.ac.uk/ris/systems/altmetric>)

Reliability and validity

How reliable and complete are the findings? The analytical framework is based on a small set of case studies, and the validation study used selections of papers that belong to the top 10% most cited publications in a category and published in the same year. There is no objective label that classifies a paper as a breakthrough; a 'breakout' is linked to the fact that the paper is regarded as such by one of the algorithms. There might be more and other characteristic patterns related to breakthroughs 'hidden' in the citation profiles. Due to the absence of an objective labelling of publications as 'breakthrough' or 'non-breakthrough', precision and recall analysis could not be executed. The impression that emerges from the validation study is that the algorithms do a fair job, but the issue of reliability and completeness needs to be addressed in follow-up studies.

Discovery date The moment a discovery is disseminated to the 'public' - in order words when it becomes known outside the group of directly involved scientists, students, technicians and supervisors - usually marks the point in time that information is shared widely amongst peers and scholars by publishing the discovery in the open scholarly literature. This $t=0$ marker should ideally be clearly identifiable in databases and identical when combining information from different sources. However, publication dates of scholarly publications are usually only an approximation of $t=0$, mainly because of the key publication outlets: peer-reviewed scholarly journals and conference proceedings. These outlets differ in publication schedules (weekly, monthly, bi-monthly, quarterly, semi-annual, annual, or irregular), which make it difficult to use a time unit of less than 1 year. Choices have to be made to determine the publication's approximate submission date. However, the results can be no more than a 'best guess' of the moment the information became public. A reasonably reliable - albeit not perfect - date is the point when a manuscript is submitted to the publisher. Research publications usually only mention the publication date, and not the date when the publication was submitted for publication. Unfortunately this submission date is in general not available in the bibliographic databases. For small samples, individual publications can be inspected to retrieve the submission date, as this date is usually mentioned in the typeset version of the manuscript.

Do papers with a large number of authors have a disproportional influence on outcomes? It is known that in some science fields papers are published with hundreds or even thousands of authors. The number of authors of a paper itself does not influence the outcomes of the algorithms, as it is not a parameter in any of the algorithms. The algorithm that might be influenced is the Discoverers-Intra-group Impact algorithm (DII) as it focuses on citation-related publications from a group, or major subgroup, of the authors of the breakout publication. In practice no issues were found. This might be due to the fact that large research groups suffer from 'inertia' that prevents them

to publish several related papers within a short time interval. Furthermore the constraint, built into the DII algorithm, that at least 66% of the original authors should also be author of a follow-up publication seems to protect the algorithm against a disproportional influence on the outcomes by papers with many authors.

Is there an issue with multidisciplinary in combination with the CWTS document clustering method? Multitidisciplinarity in this thesis is defined as the result of the dispersion of knowledge among different WoS Subject Categories or among different CWTS document clusters at the meso-level. The Cross Disciplinary Impact (CDI) algorithm therefore comes in two flavours, one for WoS Subject Categories (CDI_SC) and the other for CWTS document clusters (CDI_DC). For CDI_DC the threshold values for the parameters are such that many more documents are selected than with the CDI_SC algorithm which uses WoS subject categories when applied to the same data set; in the long run this differences between the two algorithms greatly disappears. The conclusion emerges that CWTS document clusters should not be used to study cross-disciplinary impact of scholarly publications in the short term, when multidisciplinary is defined as above.

7.2.2 Options for improvements and further research

The research presented in this thesis succeeded, as mentioned before, in constructing an analytical framework that enables the early-stage identification of papers that have an above average impact on science. Although the validation study shows convincing results there still is room to improve and enhance the analytical framework and the comprising algorithms comprising. Applying the algorithms to vast datasets results in document collections exclusively consisting of 'breakout' publications; approximately 0.3% of the publications of type 'article' and 'letter' covered in the WoS are seen as a breakout publication. The ability to create such breakout-concentrated data sets opens up opportunities for further research on the very nature on the evolution of science and, more specifically, on the production and role of breakthrough. The following directions for future research are foreseen (1) finding the operational limits of the algorithms, (2) widening the applicability of the algorithms, (3) analyzing data sets consisting only of breakout publications, and (4) modelling the dynamics of science.

Finding the operational limits of the algorithms The breakout-detection algorithms have been constructed on the basis of the information from a small set of case studies. To improve the knowledge of the general performance and the operational limits of the algorithms by performing a sensitivity analysis of the algorithms to determine the general operational limits. The impact of alleviating the top 10% criterion on the performance of the algorithms needs to be investigated. Precision and recall analysis should lead to better

insights in the performance characteristics of the algorithms and facilitate analyses of false positives and false negatives in order to identify the causes of erratic behaviour. A study to find out if, and if so under which conditions, reducing the time-window of the algorithms to less than two years is feasible. When shorter time windows are possible the algorithms could then not only be used for analytical purposes on historical data – ‘back casting’, but would also allow nowcasting and real-time monitoring.

Widening the applicability of the algorithms. Although the algorithms are constructed with early-stage identification of breakout papers in mind their applicability can be extended in other related directions. This is possible because what the algorithms succeed in doing is identifying papers that have an above average impact on science. Since finishing this study the algorithms have already been adapted in such a way that they can be used as a ‘sliding’ sensor¹⁰³ across the citation history of a paper this makes it possible to analyse the dynamic breakout behaviour of a publication over time. Another widening of the application domain of the algorithms is combining with sleeping beauty algorithms (van Raan, 2004, 2015) to answer questions such as “Do sleeping beauties classify as breakout papers?” and “Is the moment such a sleeping beauty awakes the moment it shows breakout behaviour?”

Analysing data sets consisting of only breakout publications. The application of the breakout algorithms facilitates the generation of datasets that consist exclusively of breakout publications. Such a dataset, that is cleaned from *noise*, should provide a clearer and more precise view on the core of scientific evolution. Does the hypothesis “A period of ‘normal’ science follows a paradigm shift (revolutionary science)” hold in general, and if so “Does it hold in all fields of science?” is a question that probably can be answered using cleaned datasets in combination with the knowledge of which algorithm or algorithms identified a publication as breakout.

Using these datasets that only contain breakouts it becomes possible to analyse the occurrence frequency of breakouts and breakthroughs (and their overlap, i.e., events that are both) in different science fields over time. Comparing cleaned datasets with their not-cleaned companions should help in identifying the determinants, if any, that distinguish a breakout from an ‘ordinary’ discovery in science. Such an analysis could also shine light on factors that differentiates breakthroughs from breakouts. Hopefully can such determinants be used to develop pre-defined ‘breakthrough markers’ to identify breakthroughs.

Modelling the dynamics of science The algorithms can be used to detect and monitor changes in the science system as far as these changes are reflected in scholarly publications, and should be of help in developing adequate models that describe the evolution of science. Such a model-based approach

¹⁰³A sensor is a device that detects or measures a physical property and responds to it.

can be used to detect deviations from the forecasted evolution using a proper model. It becomes then possible to link, at large scale, models and real-time events. In this way time delays can be circumvented, nowcasting becomes possible, and effects of (governmental) stimulus programs can be monitored. In identifying and delineating emerging research themes and uncover what are generally called ‘emerging topics’ or ‘emerging technologies’ cleaned datasets could be helpful.

7.2.3 Follow-up research: ‘Towards a typology of breakouts’

This study started with the notion, as mentioned before, that there is not a well defined general-accepted objective definition for a breakthrough. Studying breakthrough discoveries in science at large is therefore not an easy time-consuming hardly-feasible activity. The goal of this research was to develop an analytical framework for the early-stage detection of discoveries with an above impact on science. The algorithms that make up this framework however can be considered to be an operationalisation of the lacking breakthrough definition. Using this framework opens up the possibility to construct datasets that consist exclusively of publications of discoveries in science that are at least expected to have an above average impact on science, and some of them are breakthroughs. Such datasets facilitate studying the *breakthrough* phenomenon at large scale in a structured way. One of the questions emerging from this study is “*What differentiates a breakout publication from an ordinary publication?*” The presented follow-up research addresses the following three questions (1) “Does organisational co-operation influence the chances of a publication to be a breakout?”, (2) “Does the size of the research team influence the chances of a discovery to become a breakout?”, and (3) “Does the breakout character of a publication also manifests itself at later stages of the citation history, and not only immediately after publication?”

Organisational co-operation and breakouts This particular follow-up study focuses on the breakout character of a publication in relation with organisational co-operation across science fields. The presence or absence of such co-operation is measured as co-authorship between authors affiliated to organisations of different organisational type. The organisations taken into account are the main research-publication producing organisations: Universities (U), Research Institutes (R), Companies (C), and Hospitals (H).

The analysis was done by applying the algorithms to the data in the CWTS in-house version of Web of Science database, and all publications of type ‘article’ and ‘letter’ from 2007–2011 were included. Resulting in a set of breakout publications, where a breakout publication is defined as a publication that is classified as a ‘breakout’ by at least one of the algorithms. In total, 38,949 breakout publications among 4,799,020 publications were identified for the period 2007–2011. The results are differentiated along two dimensions (Table 7.1). The first dimension (Organisational category) contains the

organisational type of the affiliation of the authors and relevant combinations. The second dimension consists of the categories that form the highest level of the NOWT¹⁰⁴ classification; the categories 'Language, Information and Communication', and 'Law, Arts and Humanities' were left out as for these fields less than 15 breakout publications were found. The overall share of breakout publications is 0.8%.

Across all fields it is found that university staff authored or co-authored between 87% (Engineering Sciences) and 98% (Social and Behavioural Sciences) of the breakout publications. In general the combinations 'U+R' and 'U+C' (for these abbreviations see Table 7.1) produce a significant share of breakout publications. For 'Medical and Life Sciences', not surprisingly, in 18% of the breakouts there is a hospital affiliation for at least one of the authors. The journals in the 'multidisciplinary journals' category (this category is dominated by top journals such as Nature, Science, and PNAS¹⁰⁵) have by far the highest numbers of breakout publications, and the share of breakout publications for these journals is much higher than the overall average share of 0.8% we observed. Table 7.2 contains descriptive statistics for Nature, Science, and PNAS. Based on the share of breakout publications, all three journals are ranked in the top 50 of journals with the highest share of breakout publications. Table 7.3 presents the relative performance in producing breakout publications for different organisational categories. This relative performance is defined as the share of breakout publications for a certain combination of 'organisational category' and 'NOWT category' compared to the corresponding share in the total set of publications. In this table a '+' indicates that breakouts are overrepresented, and a '-' that they are underrepresented. Publications in the 'Engineering Sciences' from research-institute-affiliated authors, and publications from company-affiliated authors published in a multidisciplinary journal have a higher probability of being a breakout publication. In general, publications co-authored by authors of different types of organisations produce more breakouts than expected. An exception to this is 'Engineering Sciences', where only publications with authors from the combination 'U + R' have this property.

Does the size of the research team influence the chances of a discovery to become a breakout? In this second study the focus is on the question "Is the size of the research team different for breakout and non-breakout publications?" We use the team of authors as a proxy for the size of the research team. For the papers from 2007-2011 the algorithms were used to find out if the average size of the author teams for breakout papers differs from that for papers in general. Figure 7.1 shows the distribution of the number of papers vs. the size of the author team. This figure shows that the distribution is skewed. For 'all' papers the distribution tops at 3 authors, and for the breakout papers at 4 authors. The differences between the distributions are

¹⁰⁴Netherlands Observatory of Science and Technology (NOWT).

¹⁰⁵Proceedings of the National Academy of Sciences of the United States of America

Table 7.1: Distribution of breakout papers (articles, letters) across NOWT categories (2007–2011)^a

	Medical and Life Sciences	Natural Sciences	Engineering Sciences	Social and Behavioural Sciences	Multidisciplinary journals ^b
Total number of publications	2,362,512	2,066,802	508,093	360,443	75,047
Number of breakout publications detected	24,277	9,020	466	800	5,873
<i>Share of total</i>	<i>1.0%</i>	<i>0.4%</i>	<i>0.1%</i>	<i>0.2%</i>	<i>7.8%</i>
<i>Organisational category</i>					
University (U)	56.5%	69.8%	69.1%	83.6%	56.5%
Research institute (R)	4.4%	6.5%	9.2%	1.4%	3.3%
Company (C)	2.2%	1.6%	4.1%	0.3%	2.0%
Hospital (H)	1.5%	0.0%	0.0%	0.3%	0.4%
U + R	13.8%	15.7%	9.9%	7.9%	21.1%
U + H	8.0%	0.4%	0.2%	3.3%	4.1%
U + C	6.9%	5.7%	7.5%	2.5%	9.2%
U + H + R	3.6%	0.1%	0.0%	0.4%	2.2%
U + H + C + R	1.6%	0.0%	0.0%	0.1%	0.8%
U + H + C	1.5%	0.1%	0.0%	0.4%	0.5%

^a Excludes publications that could not be assigned to organisational subcategories.

^b Journals assigned to this category, by Thomson Reuters, include *Nature*, *Science*, and *PNAS*.

Table 7.2: Descriptive statistics for the multidisciplinary journals Nature, Science and PNAS (articles and letters, 2007–2011)

Multidisciplinary journal	Number of publications	Number of breakout publications	Breakout publications as share of total number of publications
<i>Nature</i>	13,041	1,966	15.1%
<i>PNAS</i>	20,173	1,758	8.7%
<i>Science</i>	12,744	1,585	12.4%

Table 7.3: Distribution of breakouts: over and under representation relative to all papers (articles, letters) in the WoS database (2007–2011)

	Medical and Life Sciences	Natural Sciences	Engineering Sciences	Social and Behavioural Sciences	Multidisciplinary Journals
<i>Organisational category</i>					
University (U)	-	-	-	-	-
Research Institute (R)	-	-	+	-	-
Company (C)	-	-	-	-	+
Hospital (H)	-	-	-	-	-
U + R	+	+	+	+	+
U + H	+	+	-	+	+
U + C	+	+	-	+	+
U + H + R	+	+	-	+	+
U + H + C + R	+	+	-	+	+
U + H + C	+	+	-	+	+

also illustrated in Table 7.4, where the percentile borders for the two distributions are presented. Zooming in on the data reveals that the number of authors that contribute to a paper depends on the organisational categories and on the organisational collaboration. In Table 7.5 we present the weighted average size of the author team of a publication. On average more authors contribute to breakout papers than to non-breakout papers. On the basis of these results we conclude that particularly breakout papers with one or more authors affiliated to a company have on average a larger author team than papers with only authors affiliated to universities, research institutes or hospitals. The data also shows that collaboration between organisations of different categories leads to substantially more authors. The average size of the author team depends on the organisational collaboration, and is significantly smaller for ‘mono organisational category’ papers. For breakout paper this difference is larger. The author teams for breakout papers are between 1.4 and 2.6 times larger than for all papers.

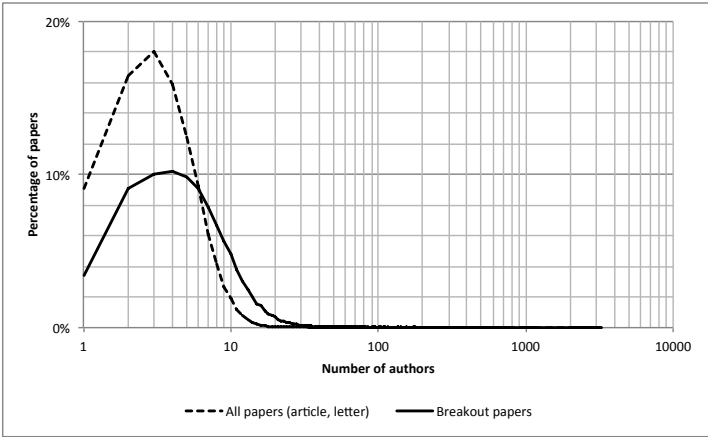


Figure 7.1: Distribution of the number of authors of a paper for ‘breakout’ and for ‘all’ papers from 2007–2011

Table 7.4: Percentile borders for the distributions of the numbers of authors for both ‘breakout’ papers and for ‘all’ papers from 2007–2011

Percentile	Breakthrough papers <i>Number of authors</i>	All papers
25%	3.2	2.0
50%	4.8	3.6
75%	9.8	5.3

Table 7.5: Weighted average number of author per paper of WoS types article and letter from 2007–2011

Breakout papers		All papers	
<i>Organisational category</i>	<i>Average number of authors per paper</i>	<i>Organisational category</i>	<i>Average number of authors per paper</i>
University (U)	6.3	University (U)	4.3
Research Institute (R)	5.8	Research Institute (R)	4.2
Company (C)	8.6	Company (C)	4.9
Hospital (H)	6.2	Hospital (H)	4.4
U + R	16.9	U + R	6.6
U + H	11.1	U + H	6.8
U + C	16.8	U + C	6.8
U + H + R	24.0	U + H + R	12.1
U + C + H + R	43.2	U + C + H + R	19.7
U + H + C	17.4	U + H + C	9.5

Table 7.6: Point in time of occurrence of the breakout character of a paper (article, letter), published in the period 1990–1994 vs. organisational collaboration

Affiliation type	Total	Years since publication a publication is identified as 'breakout publication'									
		0	1	2	3	4	5	6	7	8	9
		Number of publications									
University (U)	1,886,048	79,144	4,699	362	238	139	209	166	175	118	98
Research Organisation (R)	225,731	10,046	700	42	20	16	23	19	15	21	13
Company (C)	129,950	4,557	454	29	18	10	12	14	17	13	6
Hospital (H)	99,178	1,936	121	8	8	5	9	1	1	3	4
U + R	147,424	9,691	912	30	31	15	20	28	22	12	10
U + C	76,446	5,364	575	30	19	8	15	13	10	12	6
U + H	86,069	3,416	398	31	16	2	8	7	9	11	7
U + H + R	6,067	375	99	4	1	0	0	2	2	0	0
U + H + C	2,911	200	44	1	3	1	1	0	0	0	0
U + H + C + R	476	49	9	0	0	0	0	0	0	0	0
Total publications	2,660,300	114,778	8,011	537	354	196	297	250	251	190	144
		91.8%	6.4%	0.4%	0.3%	0.2%	0.2%	0.2%	0.2%	0.2%	0.1%

At what moment in a paper's life does the breakout character become apparent for the first time? The algorithms are developed to search for breakout characteristics from the moment of publication. In the study for this thesis the question “Does the breakout character also manifests itself at later stages of a publication's history?” was left unanswered. In order to answer this question publications of the type ‘article’ and ‘letter’ from the period 1990–1994 were selected in the WoS if the authors are affiliated with a University, a Research Organisation, a Company or a Hospital. 4.3% of these publications are marked as a breakout publication by at least one of the algorithms. The majority (98.2%) of these publications show their ‘breakout character’ within two years after publication, and 91.8% even in the first year after publication. The number of publications starting to show breakout characteristics steeply decreases after the first two years. Publications written by authors affiliated with a combination of organizations have a higher probability of being a breakout publication; publications from authors exclusively affiliated with companies or hospitals have a below average chance of being a breakout publication. Table 7.6 presents the results.

7.3 Concluding remarks

The analytical framework presented in this PhD thesis can be considered an operational, probably incomplete, definition of a breakthrough. This framework allows the generation of datasets exclusively consisting of information on scientific discoveries that are expected to have an above average impact on the evolution of science. Such clean datasets can be used for, large scale, analysis of the dynamics of the science system from the perspective of high-impact discoveries.

Bibliography

- Adams, J. (2005). Early citation counts correlate with accumulated impact. *Scientometrics*, 63:567–581. 10.1007/s11192-005-0228-9. 43, 118, 164
- Ahuja, G. and Lampert, C. M. (2001). Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal*, 22(6-7):521–543. 26
- Andersen, H., Barker, P., and Chen, X. (2006). *The cognitive structure of scientific revolutions*. Cambridge University Press, New York, NY, USA. 6, 10, 66, 92
- Andersen, P. D. and Borup, M. (2009). Foresight and strategy in national research councils and research programmes. *Technology Analysis & Strategic Management*, 21(8):917–932. 43
- Arbesman, S. (2010). Quantifying the ease of scientific discovery. *Scientometrics*, 86(2):245–250. 26, 66
- Baba, M., Tanaka, H., De Clercq, E., Pauwels, R., Balzarini, J., Schols, D., Nakashima, H., Perno, C. F., Walker, R. T., and Miyasaka, T. (1989). Highly specific-inhibition of human immunodeficiency virus type-1 by a novel 6-substituted acyclovir derivative. *Biochemical and Biophysical Research Communications*, 165(3):1375–1381. 30, 31
- Balconi, M., Breschi, S., and Lissoni, F. (2004). Networks of inventors and the role of academia: an exploration of italian patent data. *Research Policy*, 33(1):127–145. 28
- Ball, P. (2004). *Critical Mass - How One Thing Leads to Another*. Arrow Books, London. 9, 103
- Bañuls, V. A. and Salmeron, J. L. (2008). Foresighting key areas in the information technology industry. *Technovation*, 28(3):103–111. 43

- Baumgartner, S. E. and Leydesdorff, L. (2014). Group-based trajectory modeling (gbtm) of citations in scholarly literature: Dynamic qualities of “transient” and “sticky knowledge claims”. *Journal of the Association for Information Science and Technology*, 65(4):797–811. 15
- Becattini, F., Chatterjee, A., Fortunato, S., Pan, R., Parolo, P., and Mitrovic, M. (2014). *The Nobel Prize* (<http://scitation.aip.org/content/aip/magazine/physicstoday/news/10.1063/pt.5.2012>). 8
- Bettencourt, L. M. A., Kaiser, D., Kaur, J., Castillo-Chávez, C., and Wojick, D. (2008). Population modeling of the emergence and development of scientific fields. *Scientometrics*, 75(3):495–518. 13, 26, 43
- Bettencourt, L. M. A. and Kaiser, D. I. (2011). General critical properties of the dynamics of scientific discovery. Technical report, Office of Scientific and Technical Information - U.S. Department of Energy. 10, 13, 100
- Bettencourt, L. M. A. and Kaiser, D. I. (2015). Formation of scientific fields as a universal topological transition. *arXiv*, 1504.00319(v1):1–8. 10, 100, 103
- Bettencourt, L. M. A., Kaiser, D. I., and Kaur, J. (2009). Scientific discovery and topological transitions in collaboration networks. *Journal of Informetrics*, 3(3):210–221. Science of Science: Conceptualizations and Models of Science. 10, 14, 100, 101, 103, 164
- Boehm, H. P., Clauss, A., Fischer, G. O., and Hofmann, U. (1962). Das adsorptionsverhalten sehr dünner kohlenstoff-folien. *Zeitschrift für anorganische und allgemeine Chemie*, 316(3-4):119–127. 42, 60, 61
- Bonaccorsi, A. (2008). Search regimes and the industrial dynamics of science. *Minerva*, 46:285–315. 10, 100
- Bonaccorsi, A. (2010). New forms of complementarity in science. *Minerva*, 48:355–387. 10, 100
- Bonaccorsi, A. and Vargas, J. (2010). Proliferation dynamics in new sciences. *Research Policy*, 39(8):1034–1050. 100
- Brannigan, A. and Wanner, R. A. (1983). Historical distributions of multiple discoveries and theories of scientific change. *Social Studies of Science*, 13(3):417–435. 6
- Breiner, S., Cuhls, K., and Grupp, H. (1994). Technology foresight using a delphi approach - a japanese-german cooperation. *R & D management*, 24(2):141–153. 26, 44, 66
- Broadus, R. N. (1987). Toward a definition of ‘bibliometrics’. *Scientometrics*, 12(5-6):373–379. 7

- Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology*, 110(2):349–399. 12
- Chandy, R. K. and Tellis, G. J. (2000). The incumbent's curse? incumbency, size, and radical product innovation. *Journal of Marketing*, 64(3):1 – 17. 26
- Chen, C. (2012). Predictive effects of structural variation on citation counts. *Journal of the American Society for Information Science and Technology*, 63(3):431–449. 27
- Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., and Pellegrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics*, 3(3):191–209. 10, 14, 26, 44, 66, 100
- Ciechanover, A. (2006). The ubiquitin proteolytic system - from an idea to the patient bed. In *Proc Am Thorac Soc*, volume 3, pages 21–31. 19
- Ciechanover, A. (2009). Tracing the history of the ubiquitin proteolytic system: The pioneering article. *Biochemical and Biophysical Research Communications*, 387(1):1 – 10. 8, 69, 72
- Ciechanover, A. (2014). Personal communication (unpublished). 74, 81, 82
- Ciechanover, A., Elias, S., Heller, H., Ferber, S., and Hershko, A. (1980a). Characterization of the heat-stable polypeptide of the atp-dependent proteolytic system from reticulocytes. *Journal of Biological Chemistry*, 255(16):7525–7528. 73, 77, 79, 81, 84
- Ciechanover, A., Heller, H., Elias, S., Haas, A. L., and Hershko, A. (1980b). ATP-dependent conjugation of reticulocyte proteins with the polypeptide required for protein degradation. *Proceedings of the National Academy of Sciences*, 77(3):1365–1368. viii, 72, 73, 77, 78, 79, 80, 81, 84, 88
- Ciechanover, A., Hod, Y., and Hershko, A. (1978). A heat-stable polypeptide component of an ATP-dependent proteolytic system from reticulocytes. *Biochemical and Biophysical Research Communications*, 81(4):1100–1105. viii, 18, 72, 73, 78, 79, 80, 81, 84, 88
- Cintron-Arias, A., Bettencourt, L. M. A., Kaiser, D. I., and Castillo-Chávez, C. (2005). On the transmission dynamics of knowledge. Technical report, Office of Scientific and Technical Information - U.S. Department of Energy. 10, 100
- Coates, J. F., Mahaffie, J. B., and Hines, A. (1994). Technological forecasting: 1970–1993. *Technological Forecasting and Social Change*, 47(1):23–33. 44
- Coates, V., Farooque, M., Klavans, R., Lapid, K., Linstone, H. A., Pistorius, C., and Porter, A. L. (2001). On the future of technological forecasting. *Technological Forecasting and Social Change*, 67(1):1–17. 44

- Cockburn, I. M. and Henderson, R. M. (1998). Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery. *Journal of Industrial Economics*, 46(2):157-182. 28
- Cokol, M., Ozbay, F., and Rodriguez-Esteban, R. (2008). Retraction rates are on the rise. *EMBO Reports*, 9(1):2-2. 10, 122, 165
- Crane, D. (1972). *Invisible colleges: diffusion of knowledge in scientific communities*. The University of Chicago Press, Chicago. 7
- Cyranoski, D. (2004). Korea's stem-cell stars dogged by suspicion of ethical breach. *Nature*, 429(6987):3-3. 10
- Daim, T. U., Rueda, G., Martin, H., and Gerdtsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, 73(8):981-1012. Tech Mining: Exploiting Science and Technology Information Resources. 44
- Darden, L. and Maull, N. (1977). Interfield theories. *Philosophy of Science*, 44(1):43-64. 68, 106
- Davis, P. M. (2012). The persistence of error: a study of retracted articles on the internet and in personal libraries. *Journal of the Medical Library Association : JMLA*, 100(3):184-189. 122, 165
- De Clercq, E. (2009a). Antiviral drug discovery: Ten more compounds, and ten more stories (part b). *Medicinal Research Reviews*, 29(4):571-610. vii, 28, 31
- De Clercq, E. (2009b). The history of antiretrovirals: key discoveries over the past 25 years. *Reviews in Medical Virology*, 19(5):287-299. 30, 38
- De Clercq, E. (2013). Personal communication (unpublished). 34
- Dmitriyeva, O., Cantwell, R., McConnell, M., and Moddel, G. (2012). Origin of excess heat generated during loading Pd-impregnated alumina powder with deuterium and hydrogen. *Thermochimica Acta*, 543:260 - 266. 10, 122
- Epstein, D., Elias-Bishko, S., and Hershko, A. (1975). Requirement for protein synthesis in the regulation of protein breakdown in cultured hepatoma cells. *Biochemistry*, 14(23):5199-5204. 79
- Esté, J. A. and Telenti, A. (2007). HIV entry inhibitors. *The Lancet*, 370(9581):81-88. 31
- FDA (2004). *Innovation or Stagnation? Challenge and Opportunity on the Critical Path to New Medical Products*. U.S. Department of Health and Human Services - Food and Drug Administration (FDA). 28
- Fleischmann, M. and Pons, S. (1989). Electrochemically induced nuclear-fusion of deuterium. *Journal of electroanalytical chemistry*, 261(2A):301-308. 10, 122

- Frenken, J. W. M. (2013). Personal communication (unpublished). 43, 46, 53, 56, 61
- Furukawa, R. and Goto, A. (2006). The role of corporate scientists in innovation. *RESEARCH POLICY*, 35(1):24–36. 28
- Geim, A. K. and Novoselov, K. S. (2007). The rise of graphene. *Nature materials*, 6(3):183–91. 45
- Gelinas, R. J. and Roberts, R. J. (1977). One predominant 5'-undecanucleotide in adenovirus 2 late messenger RNAs. *Cell*, 11:533–544. 90
- Gerstein, M. and Douglas, S. (2007). RNAi development. *PLoS Computational Biology*, 3(4):e80–775. 14, 101
- Globe, S., Levy, G. W., and Schwartz, C. M. (1973). Science, technology and innovation. NSF-study NSF-C-667, Battelle Memorial Inst., Columbus, Ohio, Columbus Labs. 1
- Green, S. J. and Brendsel, J. (2008). Letter to the editor: Key discoveries often originate with lone researchers. *Nature*, 456:315. 12
- Grupp, H., editor (1992). *Dynamics of Science-Based Innovation*. Springer-Verlag GmbH, Heidelberg. 29
- Grupp, H. and Schmoch, U. (1992). *Dynamics of Science-Based Innovation*, chapter 9 At the crossroads in laser medicine and polyimide chemistry: patent assessment of the expansion of knowledge, pages 269–301. Springer-Verlag. 1, 46
- Guimerà, R., Uzzi, B., Spiro, J., and Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308(5722):697–702. 13, 68, 80
- Hand, D. J. (2009). Mining the past to determine the future: Problems and possibilities. *International Journal of Forecasting*, 25(3):441–451. Special Section: Time Series Monitoring. 44
- Hazuda, D. J., Anthony, N. J., Jolly, R. P., Gomez, S. M., Wai, J. S., Zhuang, L., Fisher, T. E., Embrey, M., James P. Guare, J., Egbertson, M. S., Vacca, J. P., Huff, J. R., Felock, P. J., Witmer, M. V., Stillmock, K. A., Danovich, R., Grobler, J., Miller, M. D., Espeseth, A. S., Jin, L., Chen, I.-W., Lin, J. H., Kassahun, K., Ellis, J. D., Wong, B. K., Xu, W., Pearson, P. G., Schleif, W. A., Cortese, R., Emini, E., Summa, V., Holloway, M. K., and Young, S. D. (2004a). A naphthyridine carboxamide provides evidence for discordant resistance between mechanistically identical inhibitors of hiv-1 integrase. *Proceedings of the National Academy of Sciences*, 101(31):11233–11238. 32, 33, 37
- Hazuda, D. J., Felock, P., Witmer, M., Wolfe, A., Stillmock, K., Grobler, J. A., Espeseth, A., Gabryelski, L., Schleif, W., Blau, C., and Miller, M. D. (2000). Inhibitors of strand transfer that prevent integration and inhibit hiv-1 replication in cells. *Science*, 287(5453):646–650. 29, 30, 31, 34, 37, 107

- Hazuda, D. J., Young, S. D., Cuare, J. P., Anthony, N. J., Gomez, R. P., Wai, J. S., Vacca, J. P., Handt, L., Motzel, S. L., Klein, H. J., Dornadula, G., Danovich, R. M., Witmer, M. V., Wilson, K. A. A., Tussey, L., Schleif, W. A., Gabryelski, L. S., Jin, L., Miller, M. D., and Casimiro, D. R. (2004b). Integrase inhibitors and cellular immunity suppress retroviral replication in rhesus macaques. *Science*, 305(5683):528-532. 3, 32, 33, 37
- Heilbron, J. L. (1972). IIT research institute illinois institute of technology research institute - technology in retrospect and critical events in science. *Isis*, 63(1):115. 1
- Henderson, R. M. and Clark, K. B. (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative Science Quarterly*, 35(1):9 - 30. 26
- Hershko, A. (1974). Regulation of gene expression in eukaryotic cells. In Harris, M. and Thompson, B., editors, *Fogarty International Center Proceedings*, number 25, pages 85-96, U.S. Printing Office, Washington, D.C. viii, 79
- Hershko, A. and Ciechanover, A. (1982). Mechanisms of intracellular protein breakdown. *Annual Review of Biochemistry*, 51(1):335-364. PMID: 6287917. 81
- Hershko, A., Ciechanover, A., Heller, H., Haas, A. L., and Rose, I. A. (1980). Proposed role of ATP in protein breakdown: conjugation of protein with multiple chains of the polypeptide of ATP-dependent proteolysis. *Proceedings of the National Academy of Sciences*, 77(4):1783-1786. viii, 72, 73, 77, 78, 79, 80, 81, 84, 88
- Hershko, A., Ciechanover, A., and Rose, I. A. (1979). Resolution of the ATP-dependent proteolytic system from reticulocytes: a component that interacts with ATP. *Proceedings of the National Academy of Sciences*, 76(7):3107-3110. 79, 80, 81
- Hershko, A., Heller, H., Eytan, E., Kaklij, G., and Rose, I. A. (1984). Role of the alpha-amino group of protein in ubiquitin-mediated protein breakdown. *Proceedings of the National Academy of Sciences*, 81(22):7021-7025. 77
- Hershko, A., Heller, H., Ganoth, D., and Ciechanover, A. (1978). Mode of degradation of abnormal globin chains in rabbit reticulocytes. In Segal, H. L. and Doyle, D. J., editors, *Proceedings of a Symposium on Protein Turnover and Lysosome Function*, pages 149-169, New York. Academic Press. viii, 79
- Hershko, A. and Tomkins, G. M. (1971). Studies on the degradation of tyrosine aminotransferase in hepatoma cells in culture: Influence of the composition of the medium and adenosine triphosphate dependence. *Journal of Biological Chemistry*, 246(3):710-714. 79
- Hicks, D. (2000). 360 degree linkage analysis. *Research Evaluation*, 9(2):133-143. 30

- Hofstadter, D. R. and Sander, E. (2013). *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. Basic Books, New York. 9
- Hollingsworth, J. R. (2008). Scientific discoveries: An institutionalist and path-dependent perspective. In Hannaway, C., editor, *Biomedicine in the Twentieth Century: Practices, Policies, and Politics*, volume 72 of *Biomedical and Health Research*, pages 317–353. National Institutes of Health, Bethesda, MD. 9, 26, 27, 45, 47, 67, 91, 102, 161, 166
- IIT, R. I. (1968). Technology in retrospect and critical events in science. vol. 1. Technical report, IIT Research Institute; National Science Foundation. 1
- IIT, R. I. (1969). Technology in retrospect and critical events in science. vol. 2. Technical report, IIT Research Institute; National Science Foundation. 1
- Ioannnidis, J. P. A., Boyack, K. W., Small, H., Sorenson, A. A., and Klavans, R. (2014). Is your most cited work your best? *Nature*, 514:561–562. 84
- IOM (2009). *Conflict of Medical Research, Education, and Practice*. IOM - Institute of Medicine, Washington, DC, USA. 28
- Isenson, R. S. (1969). Project hindsight (final report). Technical Report AD495905, Office of the Director of Defense Research Engineering, Washington, DC, 20301. 1, 28, 46
- Isnard, C. A. and Zeeman, E. C. (1976). Some models from catastrophe theory in the social sciences. In Collins, L., editor, *The Use of models in the social sciences*, International Behavioral and Social Sciences Ser.: Classics from the Tavistock Press, pages 44–100. Tavistock Publications. 91, 92
- Jaffe, A. B. and de Rassenfosse, G. (2016). Patent citation data in social science research: Overview and best practices. Working Paper 21868, National Bureau of Economic Research. 22
- Jewkes, J., Sawers, D., and Stillerman, R. (1958). *The Sources of Invention*. MacMillan, London. 1, 28, 46
- Jones, B. F., Wuchty, S., and Uzzi, B. (2008). Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, 322(5905):1259–1262. 13, 68, 80
- Julius, M., Berkoff, C. E., Strack, A. E., Krasovec, F., and Bender, A. D. (1977). A very early warning system for the rapid identification and transfer of new technology. *Journal of the American Society for Information Science*, 28(3):170–174. 14, 26, 44, 66
- Koshland, D. E. (2007). The cha-cha-cha theory of scientific discovery. *Science*, 317(5839):761–762. vii, 8, 10, 11, 12, 27, 44, 45, 67, 90, 91, 102, 164, 166
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. The University of Chicago Press, Chicago. 5, 9, 27, 45, 66, 83, 91, 100, 102, 166

- Lade, S. J. and Gross, T. (2012). Early warning signals for critical transitions: A generalized modeling approach. *PLoS Comput Biol*, 8(2):e1002360. 5
- Laskowski, R., MacArthur, M., Moss, D., and Thornton, J. (1993). Procheck - a program to check the stereochemical quality of protein structures. *Journal Of Applied Crystallography*, 26(2):283-291. 113
- Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, number ACM 1-59593-135-X/05/0008 in KDD '05, pages 177-187, New York, NY, USA. ACM. 14, 101
- Leydesdorff, L. and Rafols, I. (2011). Local emergence and global diffusion of research technologies: An exploration of patterns of network formation. *Journal of the American Society for Information Science and Technology*, 62(5):846-860. 44, 66
- Lievrouw, L. A. (1989). The invisible college reconsidered: Bibliometrics and the development of scientific communication theory. *Communication research*, 16(5):615-628. 7
- Lissoni, F. (2010). Academic inventors as brokers. *Research Policy*, 39(7):843-857. 28
- Lowry, O. H., Rosebrough, N. J., Farr, A. L., and Randall, R. J. (1951). Protein measurement with the folin phenol reagent. *Journal of biological chemistry*, 193(1):265-75. 74
- Luukkonen, T., Persson, O., and Sivertsen, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology, & Human Values*, 17(1):101-126. 68
- Luwel, M. and van Wijk, E. (2012). Publication delays revisited: 1998-2012. In Archambault, E., Gingras, Y., and Larivière, V., editors, *Proceedings of the 17th International Conference on Science and Technology Indicators*, volume 2, pages 569-577. 30
- Mandelbrot, B. B. (1982). *The fractal geometry of nature*. W.H. Freeman and Company, New York. 4
- Martin, B. R. (1995). Foresight in science and technology. *Technology Analysis & Strategic Management*, 7(2):139-168. 27, 42, 44, 66
- McGrew, J. F., Bilotta, J. G., and Deeney, J. M. (1999). Software team formation and decay: Extending the standard model for small groups. *Small Group Research*, 30(2):209-234. 13
- Meade, N. and Islam, T. (1998). Technological forecasting—model selection, model stability, and combining models. *Management Science*, 44(8):1115-1130. 44

- Meeussen, L., Delvaux, E., and Phalet, K. (2014). Becoming a group: Value convergence and emergent work group identities. *British Journal of Social Psychology*, 53(2):235-248. 13
- Merton, R. K. (1961). Singletons and multiples in scientific discovery: A chapter in the sociology of science. *Proceedings of the American Philosophical Society*, 105(5):470-486. 6, 102
- Merton, R. K. (1965). *On the shoulders of giants - A Shandean Postscript*. The University of Chicago Press, Chicago, Ill., USA. 5, 67
- Merton, R. K. (1968). The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56-63. 7
- Meyer, M. (2006). Are patenting scientists the better scholars?an exploratory comparison of inventor-authors with their non-inventing peers in nanoscience and technology. *Research Policy*, 35(10):1646-1662. 28
- Mishra, S., Deshmukh, S. G., and Vrat, P. (2002). Matching of technological forecasting technique to a technology. *Technological Forecasting and Social Change*, 69(1):1-27. 44
- Miyasaka, T., Tanaka, H., Baba, M., Hayakawa, H., Walker, R. T., Balzarini, J., and De Clercq, E. (1989). A novel lead for specific anti-HIV-1 agents: 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine. *Journal of Medicinal Chemistry*, 32(12):2507-2509. 30, 31
- Moed, H. F. (2005). *Citation Analysis in Research Evaluation*. Information Science and Knowledge Management. Springer Netherlands, Dordrecht. 74
- Moed, H. F., Glänzel, W., and Schmoch, U., editors (2004). *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*. Springer-Verlag New York, Inc., Secaucus, NJ, USA. 7
- Moncada, S., Palmer, R., and Higgs, E. (1991). Nitric-oxide - physiology, pathophysiology, and pharmacology. *Pharmacological Reviews*, 43(2):109-142. 113
- Narin, F. and Olivastro, D. (1998). Linkage between patents and papers: An interim EPO/US comparison. *Scientometrics*, 41:51-59. 10.1007/BF02457966. 30
- Nath, D. and Shadan, S. (2009). The ubiquitin system. *Nature*, 458(7237):421-421. 68
- Nobel Prize Committee (2004). The nobel prize in chemistry 2004 - advanced information. Technical report, Nobelprize.org. 69, 80

- Nobel Prize Committee (2010). Scientific background on the nobel prize in physics 2010 - graphene. Technical report, Class for Physics of the Royal Swedish Academy of Sciences. 18, 42, 46, 53, 60
- Novoselov, K. S., Geim, A. K., Morozov, S. V., Jiang, D., Zhang, Y., Dubonos, S. V., Grigorieva, I. V., and Firsov, A. A. (2004). Electric field effect in atomically thin carbon films. *Science*, 306(5696):666-9. 2, 4, 18, 41, 42, 60, 62, 108
- Noyons, E. C. M., van Raan, A. F. J., Grupp, H., and Schmoch, U. (1994). Exploring the science and technology interface - inventor author relations in laser medicine research. *Research Policy*, 23(4):443-457. 28
- Ogburn, W. F. and Thomas, D. (1922). Are inventions inevitable? a note on social evolution. *Political Science Quarterly*, 37(1):83-98. 6, 102
- Padhi, B. (2014). Comment on 'The Top 100 Papers - Nature explores the most-cited research of all time'. 112
- Pauwels, R., Andries, K., Desmyter, J., Schols, D., Kukla, M. J., Breslin, H. J., Raeymaeckers, A., Gelder, J. V., Woestenborghs, R., Heykants, J., Schellekens, K., Janssen, M. A. C., De Clercq, E., and Janssen, P. A. J. (1990). Potent and selective inhibition of hiv-1 replication in vitro by a novel series of tibo derivatives. *Nature*, 343(6257):470-474. 30, 31, 34
- Pendlebury, D. A. (2015). Private communication. 104
- Perla, R. J. and Carifio, J. (2005). The nature of scientific revolutions from the vantage point of chaos theory. *Science & Education*, 14:263-290. 10, 100
- Podolny, J. M. and Stuart, T. E. (1995). A role-based ecology of technological-change. *American Journal of SociologY*, 100(5):1224-1260. 26
- Ponomarev, I. V., Lawton, B. K., Williams, D. E., and Schnell, J. D. (2014a). Breakthrough paper indicator 2.0: can geographical diversity and interdisciplinarity improve the accuracy of outstanding papers prediction? *Scientometrics*, 100(3):755-765. 66, 104
- Ponomarev, I. V., Williams, D., Lawton, B., Cross, D. H., Seger, Y., Schnell, J., and Haak, L. (2012). Breakthrough paper indicator: early detection and measurement of ground-breaking research. In Jeffery, K. G. and Dvořák, J., editors, *E- Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems*, pages 295-304. 15, 27, 34, 44
- Ponomarev, I. V., Williams, D. E., Hackett, C. J., Schnell, J. D., and Haak, L. L. (2014b). Predicting highly cited papers: A method for early detection of candidate breakthroughs. *Technological Forecasting and Social Change*, 81(January 2014):49-55. 15, 44, 66, 104, 161

- Porter, A. L. and Rafols, I. (2009). Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics*, 81:719-745. 34
- Price, Derek J. de Solla Price. (1963). *Little science, big science*. Columbia University Press. 6, 7, 11, 13, 102
- Price, Derek J. de Solla Price. (1965). Networks of scientific papers. *Science*, 149(3683):510-515. 7
- Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, 25:358-359. 7
- Redner, S. (2005). Citation Statistics from 110 Years of Physical Review. *Physics Today*, 58(6):49-54. 11, 15, 27, 77, 83, 103, 104, 114, 123, 128, 130, 161
- Roberts, N., Martin, J., Kinchington, D., Broadhurst, A., Craig, J., Duncan, I., Galpin, S., Handa, B., Kay, J., Krohn, A., Lambert, R., Merrett, J., Mills, J., Parkes, K., Redshaw, S., Ritchie, A., Taylor, D., Thomas, G., and Machin, P. (1990). Rational design of peptide-based hiv proteinase inhibitors. *Science*, 248(4953):358-361. 31, 34
- Roberts, R. J. (1993). Nobel lecture: Split genes and RNA splicing. Website Nobel Prize Committee. 90
- Rose, I. A., Warms, J. V., and Hershko, A. (1979). A high molecular weight protease in liver cytosol. *Journal of Biological Chemistry*, 254(17):8135-8. 77
- Rosenkopf, L. and Nerkar, A. (2001). Beyond local search: Boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal JOURNAL*, 22(4):287-306. 26
- Scheffer, M. (2009). *Critical Transitions in Nature and Society*. Princeton Studies in Complexity. Princeton University Press, Princeton, NJ, USA. 5
- Scheffer, M. (2010). Complex systems: Foreseeing tipping points. *Nature*, 467(7314):411-412. 5, 9, 103
- Scheffer, M., Bascompte, J., Brock, W. A., Brovkin, V., Carpenter, S. R., Dakos, V., Held, H., van Nes, E. H., Rietkerk, M., and Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, 461(7260):53-59. 5, 9, 103
- Schneider, J. W. and Costas, R. (2014). Identifying potential 'breakthrough' research articles using refined citation analyses: Three explorative approaches. In Noyons, E., editor, *Context Counts: Pathways to Master Big and Little Data - Proceedings of the science and technology indicators conference 2014 Leiden*, pages 551-557. 103

- Schneider, J. W. and Costas, R. (2015). Identifying potential breakthrough publications using refined citation analyses: Three related explorative approaches. pages 1-33. Available in arXiv under id 1512.01388v1. 15, 161
- Sharp, P. A. (1993). Nobel lecture: An amazing distortion in DNA induced by a methyltransferase. Website Nobel Prize Committee. 90
- Shechter, Y., Rafaeli-Eshkol, D., and Hershko, A. (1973). Influence of protease inhibitors and energy metabolism on intracellular protein breakdown in starving escherichia coli. *Biochemical and Biophysical Research Communications*, 54(4):1518-1524. 79
- Shen, Y.-C., Chang, S.-H., Lin, G. T. R., and Yu, H.-C. (2010). A hybrid selection model for emerging technology. *Technological Forecasting and Social Change*, 77(1):151-166. 44
- Simonton, D. K. (1978). Independent discovery in science and technology: A closer look at the poisson distribution. *Social Studies of Science*, 8(4):521-532. 6
- Simonton, D. K. (1979). Multiple discovery and invention: Zeitgeist, genius, or chance?. *Journal of Personality and Social Psychology*, 37(9):1603-1616. 6
- Skilton, P. (2009). Does the human capital of teams of natural science authors predict citation frequency? *Scientometrics*, 78(3):525-542. 13
- Small, H. and Klavant, R. (2011). Identifying scientific breakthroughs by combining co-citation analysis and citation context. In Noyons, E., Ngulube, P., and Leta, J., editors, *Proceedings of the 13th International Conference of the International Society for Scientometrics & Informatics (ISSI 2011)*, pages 783-793. 27
- Small, H. G. (1977). Co-citation model of a scientific specialty - longitudinal-study of collagen research. *Social Studies of Science*, 7(2):139-166. 15, 44, 66
- Small, H. G., Boyack, K. W., and Klavans, R. (2013). Identifying emerging topics by combining direct citation and co-citation. In Gorraiz, J., Schiebel, E., Gumpenberger, C., and Moed, H. F., editors, *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference 15-17 July Vienna, Austria*, volume I, pages 928-940. 15
- Steen, R. G., Casadevall, A., and Fang, F. C. (2013). Why has the number of scientific retractions increased? *PLoS ONE*, 8(7):1-9. 165
- Stokes, T. D. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Brookings Institute Press, Washington, DC, USA. 28
- Stokols, D., Hall, K. L., Taylor, B. K., and Moser, R. P. (2008). The science of team science: Overview of the field and introduction to the supplement. *American Journal of Preventive Medicine*, 35(2, Supplement):S77-S89. 80

- Sung, J. (2008). Embodied anomaly resolution in molecular genetics: A case study of RNAi. *Foundations of Science*, 13(2):177-193. 10, 100
- Tijssen, R. J. W. (2002). Science dependence of technologies: evidence from inventions and their inventors. *Research Policy*, 31(4):509-526. 28, 30
- Tijssen, R. J. W. (2010). Discarding the 'basic science/applied science' dichotomy: A knowledge utilization triangle classification system of research journals. *Journal of the American Society for Information Science and Technology*, 61(9):1842-1852. 20, 54, 93, 107
- Trajtenberg, M. (1990a). *Economic analysis of product innovation: The case of CT scanners*. Harvard University Press. 26
- Trajtenberg, M. (1990b). A penny for your quotes - patent citations and the value of innovations. *Rand Journal of Economics*, 21(1):172-187. 26
- Tu, Y.-N. and Seng, J.-L. (2012). Indices of novelty for emerging topic detection. *Information Processing & Management*, 48(2):303 - 325. 44
- Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological Bulletin*, 63(6):384-399. 13
- Tushman, M. L. and Anderson, P. (1986). Technological discontinuities and organizational environments. *Administrative Science Quarterly*, 31(3):439-465. 26
- Upham, S. P., Rosenkopf, L., and Ungar, L. (2010). Positioning knowledge: schools of thought and new knowledge creation. *Scientometrics*, 83(2):555-581. 68
- Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157):468-472. 12, 68, 80
- Uzzi, B. and Spiro, J. (2005). Collaboration and creativity: The small world problem. *American Journal of Sociology*, 111(2):447-504. 12, 14, 68, 80, 101
- van Andel, P. (1994). Anatomy of the unsought finding. serendipity: Origin, history, domains, traditions, appearances, patterns and programmability. *The British Journal for the Philosophy of Science*, 45(2):631-648. 11, 27, 45, 91
- van Noorden, R. (2011). Science publishing: The trouble with retractions. *Nature*, 478(7367):26-28. 122, 165
- van Noorden, R., Maher, B., and Nuzzo, R. (2014). The top 100 papers - nature explores the most-cited research of all time. *Nature*, 514(7524):550-553. 83, 112, 113
- van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, 59(3):467-472. 47, 122, 170

- van Raan, A. F. J. (2015). Dormitory of physical and engineering sciences: Sleeping beauties may be sleeping innovations. *PLoS ONE*, 10(e0139786):1–38. 122, 170
- Vitanov, N. and Ausloos, M. (2012). Knowledge epidemics and population dynamics models for describing idea diffusion. In Scharnhorst, A., Börner, K., and van den Besselaar, P., editors, *Models of Science Dynamics*, volume 69 of *Understanding Complex Systems*, pages 69–125. Springer Berlin / Heidelberg. 10.1007/978-3-642-23068-4_3. 10, 100
- von Bertalanffy, L. (1969). *General System Theory*. George Braziller, Inc, 227 Broadway, New York. 4
- Wallace, P. R. (1947). The band theory of graphite. *Physical Review*, 71(9):622–634. 2, 42, 60, 61
- Walsh, J. (1973). Project review: Technological innovation: New study sponsored by NSF takes socioeconomic, managerial factors into account. *Science*, 1820:846–847. 1
- Waltman, L. and van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12):2378–2392. 17, 22, 105
- Wang, D., Song, C., and Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, 342(6154):127–132. 27, 67, 83, 118, 164
- Wang, J., Veugelers, R., and Stephan, P. (2016). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. Working Paper 22180, National Bureau of Economic Research. 15
- Wang, R., Gao, Y., and Lai, L. (2000). Ligbuilder: A multi-purpose program for structure-based drug design. *Journal of Molecular Modeling*, 6(7-8):498–516. 28
- Weinberg, A. M. (1970). Scientific teams and scientific laboratories. *Daedalus*, 99(4):1056–1075. 13, 68, 80
- Whitfield, J. (2008). Collaboration: Group theory. *Nature*, 455(7214):720–723. 12
- Wild, C., Greenwell, T., and Matthews, T. (1993). A synthetic peptide from HIV-1 gp41 is a potent inhibitor of virus-mediated cell–cell fusion. *AIDS Research and Human Retroviruses*, 9(11):1051–1053. 31
- Wilkinson, K. D., Urban, M. K., and Haas, A. L. (1980). Ubiquitin is the atp-dependent proteolysis factor i of rabbit reticulocytes. *Journal of Biological Chemistry*, 255(16):7529–7532. 73, 77, 79, 81, 84

- Winnink, J. J. (2012). Searching for structural shifts in science: Graphene R&D before and after novoselov et al. (2004). In Archambault, E., Gingras, Y., and Larivière, V., editors, *Proceedings of the 17th International Conference on Science and Technology Indicators (STI 2012)*, volume 2, pages 837–846. viii, 55, 64, 90, 92
- Winnink, J. J. and Tijssen, R. J. W. (2011). R&D dynamics in the development of HIV/AIDS drugs. In Noyons, E., Ngulube, P., and Leta, J., editors, *Proceedings of the 13th International Conference of the International Society for Scientometrics & Informatics (ISSI 2011)*, pages 855–860. 90
- Winnink, J. J. and Tijssen, R. J. W. (2013). The discovery of ‘the ubiquitin-mediated proteolytic system’: an example of revolutionary science? In Gorraiz, J., Schiebel, E., Gumpenberger, C., and Moed, H. F., editors, *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference 15–17 July Vienna, Austria*, volume I, pages 583–590. 107
- Winnink, J. J. and Tijssen, R. J. W. (2014). R&D dynamics and scientific breakthroughs in HIV/AIDS drugs development: the case of integrase inhibitors. *Scientometrics*, 101(1):1–16. 105, 107
- Winnink, J. J. and Tijssen, R. J. W. (2015). Early stage identification of breakthroughs at the interface of science and technology: lessons drawn from a landmark publication. *Scientometrics*, 102(1):113–134. viii, 55, 61, 64, 105, 107, 108
- Winnink, J. J., Tijssen, R. J. W., and van Raan, A. F. J. (2013). The discovery of ‘introns’: an analysis of the science-technology interface. In Hinze, S. and Lottman, A., editors, *Translational twists and turns: science as a socio-economic endeavor - Proceedings of STI 2013 Berlin (18th International Conference on Science and Innovation Indicators)*, pages 427–438. 105
- Winnink, J. J., Tijssen, R. J. W., and van Raan, A. F. J. (2015). Theory-changing breakthroughs in science: the impact of research teamwork on scientific discoveries. *Journal of the Association for Information Science and Technology (JASIST)*, on-line first (21 Mar 2015):1–14. 105, 108
- Wooding, S. (2007). Setting the historical context for project retrosight. RAND Rand Europe working paper series WR-466-RS, RAND Europe Cambridge Westbrook Centre, Milton Road, Cambridge. CB4 1YG. UK. 1
- Worrall, J. (2003). *Thomas Kuhn*, chapter 3 Normal Science and Dogmatism, Paradigms and Progress: Kuhn ‘versus’ Popper and Lakatos, pages 65–100. *Contemporary Philosophy in Focus*. Cambridge University Press, New York, NY, USA. 92
- Wray, K. B. (2011). *Kuhn’s Evolutionary Social Epistemology*. Cambridge University Press, Cambridge, UK; New York. 5, 67

- Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039. 11, 12, 68, 80
- Yoon, B. and Park, Y. (2007). Development of new technology forecasting algorithm: Hybrid approach for morphology analysis and conjoint analysis of patent information. *IEEE transactions on engineering management*, 54(3):588–599. 44
- Young, P. (1993). Technological growth curves: A competition of forecasting models. *Technological Forecasting and Social Change*, 44(4):375–389. 44
- Zucker, L. G. and 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):12709–16. 28

Author Index

- Adams 43, 118, 164
Ahuja 26
Amaral 13, 68, 80
Andersen 6, 10, 43, 66, 92
Anderson 26
Andries 30, 31, 34
Anthony 3, 32, 33, 37
Arbesman 27, 66
Ausloos 10, 100
- Baba 30, 31
Balconi 28
Ball 9, 103
Balzarini 30, 31
Bañuls 43
Barabási 27, 67, 83, 118, 164
Barker 6, 10, 66, 92
Bascompte 5, 9, 103
Baumgartner 15
Becattini 8
Bender 14, 27, 44, 66
Berkoff 14, 27, 44, 66
Bettencourt 10, 13, 14, 27, 43, 100,
101, 103, 164
Bilotta 13
Blau 29-31, 34, 37, 107
Boehm 42, 60, 61
Bonaccorsi 10, 100
Borup 43
Boyack 15, 84
Brannigan 6
- Breiner 27, 44, 66
Brendsel 12
Breschi 28
Breslin 30, 31, 34
Broadhurst 31, 34
Broadus 7
Brock 5, 9, 103
Brovkin 5, 9, 103
Burt 13
- Cantwell 10, 122
Carifio 10, 100
Carpenter 5, 9, 103
Casadevall 165
Casimiro 3, 32, 33, 37
Castillo-Chávez 10, 13, 27, 43, 100
Chandy 26
Chang 44
Chatterjee 8
Chen 6, 10, 14, 26, 27, 32, 33, 37, 44,
66, 92, 100
Ciechanover viii, 8, 18, 19, 69, 72-74,
78-82, 84, 88
Cintron-Arias 10, 100
Clark 26
Clauss 42, 60, 61
Coates 44
Cockburn 28
Cokol 10, 122, 165
Cortese 32, 33, 37
Costas 15, 103, 161

- Craig 31, 34
 Crane 7
 Cross 15, 27, 34, 44
 Cuare 3, 32, 33, 37
 Cuhls 27, 44, 66
 Cyranoski 10
- Daim 44
 Dakos 5, 9, 103
 Danovich 3, 32, 33, 37
 Darby 28
 Darden 68, 106
 Davis 122, 165
 De Clercq vii, 28, 30, 31, 34, 38
 de Rassenfosse 22
 Deeney 13
 Delvaux 13
 Deshmukh 44
 Desmyter 30, 31, 34
 Dmitriyeva 10, 122
 Dornadula 3, 32, 33, 37
 Douglas 14, 101
 Dubonos 2, 4, 18, 41, 42, 60, 62, 108
 Duncan 31, 34
- Egbertson 32, 33, 37
 Elias viii, 72, 73, 78-81, 84, 88
 Elias-Bishko 79
 Ellis 32, 33, 37
 Embrey 32, 33, 37
 Emini 32, 33, 37
 Epstein 79
 Espeseth 29-34, 37, 107
 Esté 31
 Eytan 77
- Faloutsos 14, 101
 Fang 165
 Farooque 44
 Farr 74
 FDA 28
 Felock 29-34, 37, 107
 Ferber 73, 78, 79, 81, 84
 Firsov 2, 4, 18, 41, 42, 60, 62, 108
 Fischer 42, 60, 61
 Fisher 32, 33, 37
 Fleischmann 10, 122
- Fortunato 8
 Frenken 43, 46, 53, 56, 61
 Furukawa 28
- Gabryelski 3, 29-34, 37, 107
 Galpin 31, 34
 Ganoth viii, 79
 Gao 28
 Geim 2, 4, 18, 41, 42, 45, 60, 62, 108
 Gelder 30, 31, 34
 Gelinas 90
 Gerdtsri 44
 Gerstein 14, 101
 Globe 1
 Gomez 3, 32, 33, 37
 Goto 28
 Green 12
 Greenwell 31
 Grigorieva 2, 4, 18, 41, 42, 60, 62, 108
 Grobler 29-34, 37, 107
 Gross 5
 Grupp 1, 27, 28, 44, 46, 66
 Guimerà 13, 68, 80
- Haak 15, 27, 34, 44, 66, 104, 161
 Haas viii, 72, 73, 78-81, 84, 88
 Hackett 15, 44, 66, 104, 161
 Hall 80
 Hand 44
 Handa 31, 34
 Handt 3, 32, 33, 37
 Hayakawa 30, 31
 Hazuda 3, 29-34, 37, 107
 Heilbron 1
 Held 5, 9, 103
 Heller viii, 72, 73, 77-81, 84, 88
 Henderson 26, 28
 Hershko viii, 18, 72, 73, 77-81, 84, 88
 Heykants 30, 31, 34
 Hicks 30
 Higgs 113
 Hines 44
 Hod viii, 18, 72, 73, 78-81, 84, 88
 Hofmann 42, 60, 61
 Hofstadter 9

- Hollingsworth 9, 26, 27, 45, 47, 67,
91, 102, 161, 166
Holloway 32, 33, 37
Horowitz 10, 14, 26, 44, 66, 100
Hou 10, 14, 26, 44, 66, 100
Huff 32, 33, 37

IIT 1
Ioannnidis 84
IOM 28
Isenson 1, 28, 46
Islam 44
Isnard 91, 92

Jaffe 22
James P. Guare 32, 33, 37
Janssen 30, 31, 34
Jewkes 1, 28, 46
Jiang 2, 4, 18, 41, 42, 60, 62, 108
Jin 3, 32, 33, 37
Jolly 32, 33, 37
Jones 11-13, 68, 80
Julius 14, 27, 44, 66

Kaiser 10, 13, 14, 27, 43, 100, 101,
103, 164
Kaklij 77
Kassahun 32, 33, 37
Kaur 10, 13, 14, 27, 43, 100, 101,
103, 164
Kay 31, 34
Kinchington 31, 34
Klavans 15, 44, 84
Klavant 27
Klein 3, 32, 33, 37
Kleinberg 14, 101
Koshland vii, 8, 10-12, 27, 45, 67, 90,
91, 102, 164, 166
Krasovec 14, 27, 44, 66
Krohn 31, 34
Kuhn 5, 9, 27, 45, 66, 83, 91, 100,
102, 166
Kukla 30, 31, 34

Lade 5
Lai 28
Lambert 31, 34

Lampert 26
Lapid 44
Laskowski 113
Lawton 15, 27, 34, 44, 66, 104
Leskovec 14, 101
Levy 1
Leydesdorff 15, 44, 66
Lievrouw 7
Lin 32, 33, 37, 44
Linstone 44
Lissoni 28
Liu 10, 14, 26, 44, 66, 100
Lowry 74
Luukkonen 68
Luwel 30

MacArthur 113
Machin 31, 34
Mahaffie 44
Maher 83, 112, 113
Mandelbrot 4
Martin 27, 31, 34, 42, 44, 66
Matthews 31
Mauil 68, 106
McConnell 10, 122
McGrew 13
Meade 44
Meeussen 13
Merrett 31, 34
Merton 5-7, 67, 102
Meyer 28
Miller 3, 29-34, 37, 107
Mills 31, 34
Mishra 44
Mitrovic 8
Miyasaka 30, 31
Moddel 10, 122
Moed 74
Moncada 113
Morozov 2, 4, 18, 41, 42, 60, 62, 108
Moser 80
Moss 113
Motzel 3, 32, 33, 37
Mukherjee 12, 68, 80

Nakashima 30, 31
Narin 30

- Nath 68
 Nerkar 26
 Nobel Prize Committee 18, 42, 46, 53, 60, 69, 80
 Novoselov 2, 4, 18, 41, 42, 45, 60, 62, 108
 Noyons 28
 Nuzzo 83, 112, 113
- Ogburn 6, 102
 Olivastro 30
 Ozbay 10, 122, 165
- Padhi 112
 Palmer 113
 Pan 8
 Park 44
 Parkes 31, 34
 Parolo 8
 Pauwels 30, 31, 34
 Pearson 32, 33, 37
 Pellegrino 10, 14, 26, 44, 66, 100
 Pendlebury 104
 Perla 10, 100
 Perno 30, 31
 Persson 68
 Phalet 13
 Pistorius 44
 Podolny 26
 Ponomarev 15, 27, 34, 44, 66, 104, 161
 Pons 10, 122
 Porter 34, 44
 de Solla Price 6, 7, 11, 13, 102
 Pritchard 7
- Raeymaeckers 30, 31, 34
 Rafaeli-Eshkol 79
 Rafols 34, 44, 66
 Randall 74
 Redner 11, 15, 27, 77, 83, 103, 105, 114, 123, 128, 130, 161
 Redshaw 31, 34
 Rietkerk 5, 9, 103
 Ritchie 31, 34
 Roberts 31, 34, 90
 Rodriguez-Esteban 10, 122, 165
- Rose viii, 72, 73, 77–81, 84, 88
 Rosebrough 74
 Rosenkopf 26, 68
 Rueda 44
- Salmeron 43
 Sander 9
 Sawers 1, 28, 46
 Scheffer 5, 9, 103
 Schellekens 30, 31, 34
 Schleif 3, 29–34, 37, 107
 Schmoch 1, 28, 46
 Schneider 15, 103, 161
 Schnell 15, 27, 34, 44, 66, 104, 161
 Schols 30, 31, 34
 Schwartz 1
 Seger 15, 27, 34, 44
 Seng 44
 Shadan 68
 Sharp 90
 Shechter 79
 Shen 44
 Simonton 6
 Sivertsen 68
 Skilton 13
 Small 15, 27, 44, 66, 84
 Song 27, 67, 83, 118, 164
 Sorenson 84
 Spiro 12–14, 68, 80, 101
 Steen 165
 Stephan 15
 Stillerman 1, 28, 46
 Stillmock 29–34, 37, 107
 Stokes 28
 Stokols 80
 Strack 14, 27, 44, 66
 Stringer 12, 68, 80
 Stuart 26
 Sugihara 5, 9, 103
 Summa 32, 33, 37
 Sung 10, 100
- Tanaka 30, 31
 Taylor 31, 34, 80
 Telenti 31
 Tellis 26
 Thomas 6, 31, 34, 102

- Thornton 113
Tijssen viii, 20, 28, 30, 54, 55, 61, 64, 90, 93, 105, 107, 108
Tomkins 79
Trajtenberg 26
Tu 44
Tuckman 13
Tushman 26
Tussey 3, 32, 33, 37

Ungar 68
Upham 68
Urban 73, 78, 79, 81, 84
Uzzi 11-14, 68, 80, 101

Vacca 3, 32, 33, 37
van Aniel 11, 27, 45, 91
van Eck 17, 22, 105
van Nes 5, 9, 103
van Noorden 83, 112, 113, 122, 165
van Raan 28, 47, 105, 108, 122, 170
van Wijk 30
Vargas 100
Veugelers 15
Vitanov 10, 100
von Bertalanffy 4
Vrat 44

Wai 3, 32, 33, 37
Walker 30, 31
Wallace 2, 42, 60, 61
Walsh 1

Waltman 17, 22, 105
Wang 15, 27, 28, 67, 83, 118, 164
Wanner 6
Warms 77
Weinberg 13, 68, 80
Whitfield 12
Wild 31
Wilkinson 73, 78, 79, 81, 84
Williams 15, 27, 34, 44, 66, 104, 161
Wilson 3, 32, 33, 37
Winnink viii, 55, 61, 64, 90, 92, 105, 107, 108
Witmer 3, 29-34, 37, 107
Woestenborghs 30, 31, 34
Wojick 13, 27, 43
Wolfe 29-31, 34, 37, 107
Wong 32, 33, 37
Wooding 1
Worrall 92
Wray 5, 67
Wuchty 11-13, 68, 80

Xu 32, 33, 37

Yoon 44
Young 3, 32, 33, 37, 44
Yu 44

Zeeman 91, 92
Zhang 2, 4, 18, 41, 42, 60, 62, 108
Zhuang 32, 33, 37
Zucker 28

Samenvatting (Summary in Dutch)

Jaarlijks worden er op dit moment in de belangrijkste wetenschappelijke tijdschriften meer dan 1,5 miljoen artikelen gepubliceerd. Het vinden van juist dié publicaties welke in potentie een bovengemiddelde invloed op de wetenschap zullen hebben is vergelijkbaar met het zoeken van de spreekwoordelijke speld in de hooiberg. Het is gezien het aantal publicaties dat jaarlijks wordt gepubliceerd ondoenlijk om voor elke afzonderlijke publicatie, op een objectieve wijze, te bepalen welke naar verwachting de invloed op de wetenschap zal zijn. In dit onderzoek wordt een automatische computermethode ontwikkeld om van elke afzonderlijke wetenschappelijke publicatie in een vroeg stadium, binnen 2-3 jaar na publiceren, aan de hand van het aantal keren dat andere wetenschappers naar de publicatie verwijzen te bepalen of hetgeen gepubliceerd is als belangwekkend wordt beschouwd. Met andere woorden het doel van de in dit proefschrift ontwikkelde zoekmethode is om voor die wetenschappelijke bevindingen, de vroegste tekenen van wetenschappelijke invloed te ontdekken. Wetenschappelijke publicaties verwijzen doorgaans naar relevante informatie in andere wetenschappelijke publicaties. Dergelijke verwijzingen worden citaties genoemd en alle citaties naar één en dezelfde publicatie vormen een citatieprofiel. De zoekmethode maakt gebruik van informatiebestanden die de beschrijvende (bibliografische) informatie van wetenschappelijke artikelen bevatten en analyseert aan de hand van citatiepatronen welk effect een publicatie heeft op wetenschappelijke vooruitgang.

Ten einde het hiervoor geformuleerde doel van dit onderzoek te realiseren was het nodig om geautomatiseerde methodes te ontwikkelen welke in een vroeg stadium dié wetenschappelijke ontdekkingen kunnen vinden, waarvan verwacht wordt dat ze een grote invloed zullen hebben op de wetenschap. Vaak worden dergelijke wetenschappelijke ontdekkingen aangeduid met de term *doorbraak* en ontwikkelen ze zich op termijn als basis voor nieuwe baanbrekende technologieën. Een complicerende factor is dat het be-

grip *doorbraak* niet kan rekenen op een eenduidige en alom geaccepteerde definitie. Mede hierdoor is het zoeken naar dergelijke technologisch relevante wetenschappelijke doorbraken niet triviaal. In dit onderzoek is vanwege het ontbreken van een dergelijke geaccepteerde definitie een alternatieve benadering gekozen. Door meerdere wetenschappers zijn aanwijzingen gevonden dat in grote lijnen alle wetenschappelijke ontdekkingen het gevolg zijn van dezelfde universele basismechanismen. Op grond van deze aanwijzingen is er in dit onderzoek voor gekozen om een beperkt aantal *erkende* doorbraken in detail te analyseren en te zoeken naar kenmerkende patronen in de citatieprofielen van de bij dergelijke doorbraken behorende wetenschappelijke publicaties. Als dergelijke kenmerkende patronen zijn gevonden dan kunnen zoekmethodes worden geïmplementeerd die tevens kunnen dienen als automatische identificatieprocedures. Op grond van de analyses kunnen voor elk van de gevonden zoekmethodes de citatieparameterwaarden worden bepaald die beslissen of een publicatie al dan niet moet worden gezien als een potentiële doorbraak.

De wijze waarop wetenschap zich ontwikkelt is grillig en volgt zelden een rechtstreekse route van probleemstelling naar oplossing. Vooruitgang in de wetenschap is vaak een proces van vallen en opstaan. Periodes van stagnatie worden afgewisseld met kleine stappen voorwaarts en momenten waarop een grote stap voorwaarts —een doorbraak— wordt gemaakt. Om dit gedrag van het wetenschappelijk onderzoek te duiden onderscheidt de vermaarde wetenschapsfilosoof Thomas Kuhn twee vormen van wetenschappelijke ontwikkeling. Deze twee vormen noemt hij *normale* wetenschap en *revolutionaire* wetenschap. In normale wetenschap gaat het om het oplossen van wetenschappelijke puzzels. Bij revolutionaire wetenschap worden ontdekkingen gedaan die niet binnen het bestaande theoretisch kader van een wetenschappelijk vakgebied passen, zodat een nieuwe of aangepaste theorie noodzakelijk is om de waargenomen verschijnselen te kunnen verklaren. De noodzaak om een nieuw theoretisch kader te ontwikkelen wordt door Thomas Kuhn een *paradigmaverschuiving* genoemd. In een reflectie op wetenschappelijke ontdekkingen van enkele eeuwen en voortbordurend op de ideeën van Kuhn onderscheidt Daniel Koshland, biochemicus en gedurende de periode 1985–1995 hoofdredacteur van het prestigieuze wetenschappelijke tijdschrift *Science*, drie typen van wetenschappelijke ontdekkingen *Charge*, *Challenge* en *Chance*. Ontdekkingen van het type *Charge* zijn de meest voorkomende en behoren tot hetgeen Kuhn normale wetenschap noemt. Met andere woorden er is als het ware een wetenschappelijke puzzel opgelost met bestaande stukjes. Ontdekkingen van het type *Challenge* behoren tot hetgeen Kuhn revolutionaire wetenschap noemt. Dergelijke ontdekkingen geven aanleiding tot een nieuw of sterk aangepast theoretisch kader en zijn veel minder talrijk dan ontdekkingen van het type *Change*. Het derde type wordt gevormd door ontdekkingen die bij toeval worden gedaan en waarnaar de wetenschapper niet op zoek was; een voorbeeld is de ontdekking van penicilline door Alexander Fleming in 1928. Deze ontdekkingen zijn zeer zeldzaam en vormen derde type van Koshland —de *Chance's*.

Hoewel er, zoals hiervoor opgemerkt, geen eenduidige en alom geaccepteerde definitie van de term ‘doorbraak’ is, hebben ontdekkingen die dit predicaat krijgen wel enkele opvallende algemene eigenschappen. Zo verspreidt de kennis van een dergelijke ontdekking zich snel binnen het betrokken vakgebied en wordt er aan de ontdekking gerefereerd door veel vakgenoten. De methoden die in dit onderzoek zijn ontwikkeld kijken naar de snelheid waarmee de kennis die aan de ontdekking ten grondslag ligt zich verspreidt over de wetenschap. Hierbij wordt naar verschillende aspecten gekeken. Voor het bepalen van dergelijke kenmerkende eigenschappen van een doorbraakontdekking worden een viertal ontdekkingen, waarvan experts hebben geoordeeld dat het een *doorbraak* betrof, in detail geanalyseerd. Alle vier deze ontdekkingen hebben geleid tot technologische activiteiten, zoals blijkt uit de octrooiactiviteit die direct is te koppelen aan deze ontdekkingen. Onderzocht zijn de volgende vier wetenschappelijke ontdekkingen:

1. *Integrase Inhibitors* —een klasse van HIV/AIDS medicijnen— deze ontdekking wordt besproken in Hoofdstuk 2;
2. de ontdekking van de mogelijkheid om grafeen —lagen koolstof met een dikte van 1 atoom— te isoleren is het onderwerp van Hoofdstuk 3;
3. de ontdekking van het *ubiquitin-mediated proteolytic system* —een mechanisme dat een belangrijke rol speelt bij de afbraak van eiwitten— wordt besproken in Hoofdstuk 4;
4. de ontdekking van de *introns* —de schijnbaar overbodige loze ruimtes die deel uitmaken van genetisch materiaal en die bij het kopiëren eerst worden verwijderd om er naderhand weer te worden toegevoegd— is het onderwerp van Hoofdstuk 5.

De ontdekkingen van integrase inhibitors en grafeen zijn voorbeelden van ontdekkingen van het type Charge. Ubiquitin en intron zijn voorbeelden van Challenge ontdekkingen.

Uit de analyse van de ontdekking van de integrase inhibitors (Hoofdstuk 2) komt naar voren dat vooral onderzoekers die zowel auteur zijn van wetenschappelijke publicaties als ook vermeld zijn in octrooien die direct verwant zijn aan hun wetenschappelijke publicaties kunnen worden gezien als initiators van doorbraken. Een probleem voor een automatische analyse methode is de koppeling die tussen de wereld van de wetenschappelijke publicaties en die van de octrooien moet worden gerealiseerd. Deze koppeling vindt plaats door de namen personen in beide omgevingen te vergelijken. Auteurs van wetenschappelijke publicaties zorgen er doorgaans voor dat hun naam correct en op consistente wijze wordt vermeld op wetenschappelijke publicaties. Bij octrooien blijkt dat namen doorgaans minder consistent worden genoteerd en mogelijk op wijze die afwijkt van hetgeen in de wereld van de wetenschappelijke publicaties gebruikelijk is. De genoteerde naam van één en dezelfde persoon kan derhalve verschillend zijn in beide werelden en er is in het algemeen onvoldoende informatie beschikbaar om een automatische procedure met zekerheid te laten concluderen dat een naam verwijst naar één persoon die zowel auteur als uitvinder is. De analyses laten zien dat de methode om te zoeken naar auteurs die ook uitvinder zijn alleen praktisch uitvoerbaar

is wanneer het gaat om een beperkt aantal namen —enkele honderden tot hooguit een paar duizend. In dit onderzoek kwam tevens naar voren dat afhankelijk van het *doorbraakkarakter* van een ontdekking er een verschil is in diffusiesnelheid van de kennis over verschillende wetenschapsgebieden.

De analyse van grafeen (Hoofdstuk 3) laat zien dat de snelheid waarmee nieuwe onderzoekers en onderzoeksinstituten in de periode van 2–3 jaar na publicatie worden aangetrokken een indicatie is of een ontdekking behoort tot de ontdekkingen die een bovengemiddelde invloed heeft op nieuw wetenschappelijk onderzoek. Voor grootschalige analyses zijn de namen van auteurs goed bruikbaar, omdat auteurs, zoals ook hierboven is vermeld er doorgaans voor zorgen dat hun naam op een correcte wijze wordt genoteerd. Namen van instituten zijn enkel bruikbaar voor dataverzamelingen met niet meer dan enkele honderden of duizenden namen en wel omdat er bij het noteren van de naam van een instituut auteurs meer vrijheid hebben. Om die reden zijn er extra handelingen nodig om te beoordelen of verschillende namen ook daadwerkelijk verschillende instituten aanduiden of dat zij toch verwijzen naar hetzelfde instituut.

De analyse van ubiquitin (Hoofdstuk 4) laat zien dat bij een ontdekking die gepaard ging met een paradigmaverschuiving de kennis van de ontdekking in eerste instantie beperkt blijft tot een klein team van betrokken onderzoekers en dat pas na ruim twee jaar ook andere onderzoekers geïnteresseerd raakten. Dit onderzoek laat ook zien dat de publicatiedatum, zoals die in de bibliografische databases is opgeslagen, niet goed geschikt is om ontwikkeling van de wetenschap te analyseren op een tijdschaal met een resolutie van minder dan één jaar. Een ander effect, dat werd waargenomen, is het ontstaan van *clusters* van citerende publicaties die enerzijds de doorbraakpublicatie citeren en anderzijds ook onderlinge citatierelaties hebben. Het optreden van dit effect is vooral relevant gezien de korte tijdschaal die gebruikt wordt in de analyses, omdat binnen dat tijdsbestek de kennis voornamelijk nog geconcentreerd is bij de groep van onderzoekers die het onderzoek uitvoeren.

De ontdekking van intron (Hoofdstuk 5) laat zien dat er ongeveer 10–13 jaar na deze Challenge ontdekking een plotselinge overgang zichtbaar is waarbij de tot dan toe voornamelijk theoretische kennis migreert in de richting van meer toegepast wetenschappelijk onderzoek. De geconstateerde plotse toegenomen activiteit bleek niet te koppelen aan één specifieke ‘vervolgontdekking’, maar het resultaat te zijn van de verdere ontwikkeling van het intron-veld als geheel.

Uitgaande van de citatiepatronen die zijn gevonden is een vijftal zoekalgoritmen ontwikkeld en geïmplementeerd. De vijf algoritmen richten zich op:

1. de mate van de instroom van nieuwe actieve onderzoekers in het desbetreffende wetenschappelijke onderwerp;
2. de verspreiding van kennis afkomstig van de ontdekking over verschillende wetenschapsvelden;
3. ontdekkingen waarvan de verdere kennisontwikkeling en -verspreiding in de beginperiode voornamelijk beperkt blijft tot een klein team van

- nauw samenwerkende onderzoekers;
4. publicaties die de overgang markeren van fundamenteel wetenschappelijk onderzoek naar meer toepassingsgericht onderzoek;
 5. ontdekkingen waarin gedurende de eerste 2-3 jaar relatief veel publicaties zowel de doorbraakpublicatie citeren alsmede onderlinge citatierelaties hebben.

De vijf algoritmen hebben elk hun eigen citatieparameters, omdat ze elk een ander facet van de impact van wetenschappelijke ontdekkingen beschrijven. Bij de validatie van de methoden en de bijbehorende algoritmen blijkt dat in het algemeen wetenschappelijke publicaties worden geselecteerd die als zeer relevant worden aangemerkt omdat ze òf de basis vormen voor toegekende Nobel prijzen, dan wel voorkomen op Nature's lijst met de 100 meest geciteerde publicaties ooit. Daarnaast blijken de algoritmes publicaties te selecteren die of behoren tot de top 10% meest in overzichtsartikelen (review papers) geciteerde publicaties, of tot de top 10% meest in octrooien geciteerde wetenschappelijke publicaties, of beide. De methoden vertonen een onderlinge overlap in hun resultaten, maar ook duidelijk verschillen.

De ontwikkelde methodes en algoritmen lossen in elk geval voor een aanzienlijk deel het probleem van het ontbreken van een eenduidige alom geaccepteerde definitie van het begrip doorbraak op en maken het mogelijk om specifieke dataverzamelingen te construeren die bestaan uit ontdekkingen welke, in elk geval vlak na hun publicatie, worden gezien als belangrijk voor nieuwe ontwikkelingen in de wetenschap.

Curriculum Vitae — Jos Winnink

English version

Jos Winnink was born on April 8, 1953 in The Hague. He received his MSc in Analytical Chemistry from the University of Amsterdam. In 1978 he started his career at the Netherlands Bureau for Economic Policy Analysis (CPB); first as statistical analyst, and later as computer scientist. In 1998 he joined the Netherlands Patent Office (NPO) to become supervisor of the IT department, until in 2003 the board of NPO invited him to start a small department for patent information analysis to support the innovation policy of the Dutch government. In 2009 he became senior advisor patent information analysis. Jos Winnink authored and co-authored several reports on a variety of subjects in relation to patents and other types of industrial property, as well as emerging technologies.

At the Centre for Science and Technology Studies (CWTS) of Leiden University, he conducted research on the interface of science and technology. Main results of this research are presented in the PhD thesis. His main research interests are the early identification of potential scientific breakthroughs and the role scientific discoveries in science, and especially breakthroughs play in the development of new technologies. In this research he combines information from scholarly publications and patent publications. Jos has published several papers on his research topics and presented results at several scientific conferences.

Nederlandse versie

Jos Winnink is op 8 april 1953 geboren in Den Haag. Zijn doctoraal Analytische Scheikunde heeft hij behaald aan de Universiteit van Amsterdam. In 1978 begon zijn professionele loopbaan bij het Centraal Planbureau (CPB); eerst als statistisch medewerker en later als informaticus. Hij verruilde in 1998 het CPB voor het Bureau voor de Industriële Eigendom om hoofd van de automatiseringsafdeling te worden. Op verzoek van de toenmalige directie in

2003 heeft hij een kleine afdeling opgezet die als taak had om octrooi informatie te analyseren om beleidsrelevante informatie ten behoeve van de Nederlandse overheidsbeleid aan te leveren. Bij een reorganisatie in 2009 werd de zelfstandige afdeling voor octrooianalyse ondergebracht in een groter organisatorisch geheel en werd hij senior beleidsadviseur met als specialisatie octrooi informatie. Jos Winnink heeft als auteur en als coauteur meegewerkt aan rapporten over verschillende onderwerpen gerelateerd aan octrooien en andere intellectuele eigendomsrechten en zich ontwikkelende technologieën.

Bij het Centrum voor Wetenschaps- en Technologiestudies (CWTS) van de Universiteit Leiden heeft hij onderzoek gedaan op het raakvlak van wetenschap en technologie. De belangrijkste resultaten worden gepresenteerd in dit academisch proefschrift. Jos is vooral geïnteresseerd in het vroegtijdig herkennen van potentiële doorbraken in de wetenschap en de rol van wetenschappelijke ontdekkingen in de ontwikkeling van nieuwe technologieën. In het onderwerp dat is beschreven in dit proefschrift is informatie over wetenschappelijke publicaties gecombineerd met die over octrooipublicaties. Jos heeft verschillende artikelen over zijn onderzoek gepubliceerd en de resultaten gepresenteerd op meerdere wetenschappelijke conferenties.

This document was typeset using pdfTeX[®] and the L^AT_EX[®] environment, with the following packages:

a4wide	epstopdf	multicol
amsfonts	fancyhdr	natbib
amsmath	fncychap	placeins
amssymb	fontenc	shapepar
appendix	framed	supertabular
appendix	geometry	tabularx
array	graphicx	textcomp
authorindex	hhline	textcomp
babel	hyperref	tikz
calc	longtable	url
chngcntr	lucbmath	verbatim
cmbright	lucidabr	xcolor
crop	makeidx	
enumitem	moresize	

Ludica[®] Bright (Design foundry: Bigelow & Holmes) is the typeface used for the cover of this book

T_EXStudio[®] was used as the main type editing environment

BibDesk[®] and JabRef[®] were used to maintain the (BibT_EX[®]) bibliography database