Frank Havemann

Einführung in die Bibliometrie

Frank Havemann

Einführung in die Bibliometrie

Gesellschaft für Wissenschaftsforschung Berlin

2009

Deutsche Nationalbibliothek
CIP-Einheitstitelaufnahme
Havemann, Frank:
Einführung in die Bibliometrie /
Frank Havemann. – Berlin:
Gesellschaft für Wissenschaftsforschung, 2009
ISBN: 978-3-934682-46-7
NE: Havemann, Frank

1. Auflage 2009
Gesellschaft für Wissenschaftsforschung
c/o Institut für Bibliotheks- und
Informationswissenschaft
der Humboldt-Universität zu Berlin
Unter den Linden 6, D-10099 Berlin
`http://www.wissenschaftsforschung.de`
Redaktionsschluss: 8. April 2009
This is an Open Access e-book licensed under
the Creative Commons License BY `http:`
`//creativecommons.org/licenses/by/2.0/`

Vorwort

Will man die für ein Thema relevante wissenschaftliche Literatur auffinden, so verhelfen heute über das Web verfügbare bibliographische Datenbanken in den meisten Fachgebieten zum Erfolg. In einigen dieser Datenbanken sind neben den klassischen bibliographischen Angaben auch die in Aufsätzen zitierten Referenzen erfasst.

Es wäre eine Dummheit, diesen riesigen Schatz an Daten zur wissenschaftlichen Literatur allein für das *Information Retrieval* zu nutzen. Er ruft geradezu danach, statistisch ausgewertet zu werden. Durch Analyse dieser Daten können Erkenntnisse über die Entwicklung der Wissenschaft gewonnen werden, die sich die Wissensgesellschaft nicht entgehen lassen sollte. Wissenschaftsforschung und -soziologie können an diesem Datenmaterial Hypothesen testen. Wissenschaftspolitik kann mittels Indikatoren die Wissenschaftsentwicklung beobachten. Auch Patentdatenbanken enthalten eine Menge aufschlussreicher Informationen zur Entwicklung von Wissenschaft und Technologie.

Bibliometrie stellt die für die statistische Analyse bibliographischer und Patentdatenbanken nötigen Methoden zur Verfügung. Sie entwickelt zuverlässige Indikatoren für alle interessanten Aspekte der Wissenschaft. Von Bibliometrikern[1] entwickelte Konzepte finden auch bei der Analyse des Web Anwendung, ihre Methoden sind nicht nur auf wissenschaftliche und Patentliteratur anwendbar.

Bibliotheks- und Informationswissenschaft kann bibliometrische Erkenntnisse bei der Optimierung der Informationsversorgung nutzen, einmal direkt, weil Bibliometrie archivierte Kommunikation untersucht, und auch indirekt, indem sie in ihrer Forschung und Entwicklung bibliometrische Analysen zum Wandel der Wissenschaft berücksichtigt.

Diese kurze Einführung in die Bibliometrie wendet sich an alle diejenigen, die grundlegende Erkenntnisse und wichtige Methoden dieses interdisziplinären Forschungsgebietes kennenlernen wollen, sei es, weil sie selber Analysen durchführen wollen, oder weil sie als Rezipienten bibliometrisch gewonnener Resultate besser Grenzen und Möglichkeiten dieses Gebietes erfassen wollen. Der Text entstand nach einigen Jahren Lehre zur Bibliometrie für Studierende der Bibliotheks- und Informationswissenschaft, welche durch ihn auf das Niveau einer Vorlesung gehoben werden soll.

Besonders im letzten Jahrzehnt hat sich die Verwendung bibliometrisch gewonnener Kennziffern für die Evaluation von Forschung verbreitet. Diese Problematik wird hier ebenfalls diskutiert, so dass auch die davon Betroffenen möglicherweise Nutzen aus der Lektüre des Büchleins ziehen werden.

Schulmathematik sollte ausreichen, um auch die mathematischen Teile des Textes zu verstehen. Ich verweise jeweils auf Stellen in der Literatur, falls die Erläuterung des mathematischen Hintergrunds hier zu weit führen würde. Insgesamt stehen mathematische Formeln und Zusammenhänge nicht im Vordergrund der Darstellung, sondern die methodischen Probleme und die Ergebnisse. Darin unterscheidet sich diese Einführung von anderen Lehrbüchern und Monographien (vgl. Abschnitt *Literatur*, S. 10). Sie zielt auf empirische Bibliometrie im Gegensatz zur mathematischen oder theoretischen.

Auswahl und Einteilung des Stoffes, wie ich sie hier vornehme, sind nicht zwingend. Sie orientieren sich am Ziel, in konzentrierter Form den Lesern wesentliche Konzepte der Bibliometrie so nahezubringen, dass sie zu eigenen Untersuchungen befähigt werden und sich schnell in die aktuelle Literatur zu ihrem Forschungsproblem einlesen können. Deshalb unterteile ich den Stoff nach methodischen Gesichtspunkten, nicht nach den jeweiligen Untersuchungsgegenständen und gehe auch nicht historisch vor, wenngleich ich die Methoden nicht losgelöst von Geschichte und Anwendungen behandle.

Ich wäre froh, wenn es mir mit diesem Büchlein gelänge, deutlich werden zu lassen, dass die Bibliometrie ein interessantes und für viele Zwecke fruchtbares Forschungsfeld ist.

F. H., Berlin, 2008

[1] Wenn ich ausschließlich die männliche oder weibliche Form verwende, so dient dies der Lesbarkeit und Einfachheit. Es sind stets Personen des jeweils anderen Geschlechts mit einbezogen, sofern nicht ausdrücklich anders erwähnt.

Danksagung

Ich danke meinen Kollegen und Koautoren für viele interessante und erhellende Gespräche und Korrespondenzen über unser Fach: Manfred Bonitz, Hans-Jürgen Czerwon (der auch den Text nach Fehlern durchsah), Wolfgang Glänzel, Jochen Gläser, Michael Heinz (mit dem ich jede Woche mehrmals diskutiere), Matthias Kölbel, Sylvan Katz, Hildrun Kretschmer, Liang Liming, Irina Marshakova, Oliver Mitesser, Heinrich Parthey, Andrea Scharnhorst, Walther Umstätter, Roland Wagner-Döbler, Michel Zitt. Ebenfalls anregend waren Diskussionen mit den Hörern der im Winter 2008/2009 gehaltenen Vorlesung, die sich auf Vorformen dieses Textes stützte, und mit Teilnehmern früherer Kurse: Andrea Kaufmann, Martin de la Iglesia, Philipp Mayr, Marion Schmidt, Jakob Voß.

Nicht zuletzt danke ich allen, die an der frei zugänglichen Statistik- und Grafiksoftware `R` mitgewirkt haben, welche ich seit einigen Jahren mit großem Gewinn nutze.[2]

F. H., Berlin, März 2009

[2] http://www.r-project.org

Inhaltsverzeichnis

1 Bibliometrie als Forschungsgebiet **7**
- 1.1 Gegenstand . . . 7
- 1.2 Metadaten und Volltexte . . . 8
- 1.3 Historisches . . . 9
- 1.4 Methodisches . . . 10
- 1.5 Literatur . . . 10

2 Bibliometrische Verteilungen **13**
- 2.1 Die Lotka-Verteilung . . . 13
- 2.2 Exkurs: Gesetz von Zipf . . . 15
- 2.3 Die Bradford-Verteilung . . . 16
- 2.4 Lognormalverteilungen der Publikationsproduktivität . . . 18
- 2.5 Zitationsverteilungen . . . 19
- 2.6 Alterung und Wachstum von Literatur . . . 21

3 Bibliometrische Netzwerke **25**
- 3.1 Zitationsnetzwerke von Artikeln . . . 25
- 3.2 Zitationsnetzwerke von Journalen . . . 27
- 3.3 Exkurs: *PageRank*-Algorithmus . . . 30
- 3.4 Bibliographische Kopplung . . . 31
- 3.5 Kozitationsanalyse . . . 32
- 3.6 Exkurs: Vektorraum-Modell . . . 35
- 3.7 Koautorschaftsnetzwerke . . . 36

4 Bibliometrische Modelle **39**
- 4.1 Der Matthäus-Effekt . . . 39
- 4.2 Der Yule-Prozess . . . 40
- 4.3 Das Urnenmodell von Price . . . 42
- 4.4 Exkurs: Gesetz von Gibrat . . . 42
- 4.5 Skaleninvarianz . . . 44
- 4.6 Wachstumsmodelle . . . 44

5 Bibliometrische Indikatoren **47**
- 5.1 Produktivität . . . 47
- 5.2 Wirkung (*impact*) . . . 48
- 5.3 Kooperation . . . 51
- 5.4 Zitationsverhalten . . . 53

6 Anwendungen **55**
- 6.1 Evaluative Bibliometrie . . . 55
- 6.2 *Information Retrieval* . . . 56
- 6.3 Wissenschaftsforschung . . . 56

7 Ausblick **59**

Kapitel 1

Bibliometrie als Forschungsgebiet

1.1 Gegenstand

Bibliometrie als Forschungsgebiet, das sich mit der statistischen Analyse bibliographischer Informationen befasst, hat vor allem mit der Untersuchung des Stroms wissenschaftlicher Zeitschriftenaufsätze Bedeutung erlangt. In vielen Fachgebieten werden Forschungsergebnisse fast ausschließlich durch Aufsätze in Fachzeitschriften publik gemacht. Ihre große Zahl provozierte schon vor dem Aufkommen bibliographischer Datenbanken die statistische Untersuchung. Seitdem bibliographische Daten so aufbereitet werden, dass sie durch Maschinen gelesen und verarbeitet werden können, ist ihre massenhafte Analyse fast unvermeidlich geworden.

Dass Bibliometriker – im Gegensatz zu Forschern in vielen anderen Gebieten – ihre Untersuchungen auf vorhandene, strukturierte und leicht zugängliche Daten gründen können, kann jedoch das Interesse an ihren Ergebnissen nicht erklären. Warum durch Analyse der Publikationstätigkeit von Wissenschaftlern überhaupt Wesentliches über Wissenschaft herausgefunden werden kann, bedarf noch weiterer Begründung, die nur aus der zentralen Rolle der Veröffentlichung von Resultaten für das Funktionieren von wissenschaftlicher Forschung abgeleitet werden kann. Daher sind ein paar Bemerkungen über das Wesen von Wissenschaft, wie sie sich mit dem Beginn der Neuzeit herausgebildet hat, unumgänglich.

Wissenschaft zielt auf sichere Aussagen über die Wirklichkeit, die sich letztlich in der Praxis zu bewähren haben. Sie gewinnt durch Beobachtung und Experiment Daten, an denen ihre Aussagen und Aussagensysteme (Theorien) getestet werden können. Durch Forschung produziertes neues Wissen wird kollektiv abgesichert, indem es durch wissenschaftliche Fachgemeinschaften strengen Tests unterworfen wird. Diese kollektive Wissensproduktion kann nicht nur auf unmittelbarer mündlicher Kommunikation beruhen, durch die Schriftform gewinnt die Kommunikation an Bestimmtheit und räumlich wie zeitlich an Reichweite. Die Erfindung des Buchdrucks mit beweglichen Lettern hat der Wissenschaft der Neuzeit daher erst zu der für sie charakteristischen Dynamik verholfen, welche jetzt durch die Kommunikation über das Internet noch verstärkt wird.

Die wissenschaftlich Forschenden setzen ihre Ergebnisse durch die Veröffentlichung nicht nur der Kritik der Fachkollegen aus, sondern sichern sich durch ein vervielfältigtes Dokument auch den Ruhm, die ersten gewesen zu sein. Nach der Publikation ist das neue Wissen nicht mehr neu, jede folgende Äußerung zum Thema – insbesondere jede schriftliche – muss auf die erste Veröffentlichung Bezug nehmen. Durch Priorität gewonnenes Ansehen in der Fachgemeinschaft ermöglicht den Zugang zu Stellen an Universitäten und anderen Forschungseinrichtungen und erleichtert es, neue Forschungsmittel einzuwerben – was heute wegen steigender Forschungskosten immer wichtiger wird.

Diese, durch ein spezifisches Verhältnis von Zusammenarbeit und Wettbewerb charakterisierte Produktionsweise wissenschaftlichen Wissens, in der die Produkte nicht auf einem Markt verkauft werden, sondern allen frei zur Verfügung stehen, hat sich über die Jahrhunderte als sehr fruchtbringend erwiesen.[1] Sowohl die kollektive wie die kompetitive Seite der Produktion wissenschaftlichen Wissens verlangen nach

[1] Der berühmte US-amerikanische Wissenschaftssoziologe Robert K. Merton formulierte es so: "That crucial element of free and open communication is what I have described as the norm of ≫communism≪ in the social institution of science. . . Indeed, long before the nineteenth-century Karl Marx adopted the watchword of a fully realized communist society—≫from each according to his abilities, to each according to his needs≪—this was institutionalized practice in the communication system of science. This is not a matter of human nature, of nature-given altruism. Institutionalized arrangements have evolved to motivate scientists to contribute freely to the common wealth of knowledge according to their trained capacities, just as they can freely take from that common wealth what they need." (Merton 1988, S. 620)

der Publikation der Forschungsergebnisse. Was nicht publiziert ist, existiert für die Wissenschaft nicht.

Unmittelbar wirtschaftlich verwertbare Forschungsergebnisse werden oft nicht (oder nicht sofort) als Aufsätze in Fachzeitschriften publiziert, sondern als Patentschriften. Patentgesetzgebung zielt gerade darauf, durch den (zeitlich befristeten) Patentschutz für die Forscher oder Erfinder einen Anreiz zur Offenlegung ihrer Erfindungen zu schaffen, welche dadurch auch für andere als die Urheber nutzbar werden. Die Patentliteratur ist heute ebenfalls in Datenbanken recherchierbar und wird mit bibliometrischen Methoden analysiert.

Originalmitteilungen von Forschungsergebnissen in Form von Aufsätzen (*papers*) in Zeitschriften (und auch in Sammelbänden) haben eine typische Struktur, die sich aus ihrer Funktion in der Wissenschaftskommunikation ableiten lässt. Zu Beginn muss mindestens ein Forschungsproblem benannt werden, um dessen Lösung es in dem Aufsatz geht. Dann werden die von den Autoren verwendeten Methoden dargestellt, sowie die mit ihnen gewonnenen Resultate. Abschließend folgt eine bewertende Diskussion der Methoden und Ergebnisse, in der oft auf Forschungsprobleme hingewiesen wird, deren Bearbeitung durch die Ergebnisse nötig oder möglich geworden ist.

Durch die Darstellung von Problem und Methoden wird die Angemessenheit der Methoden dem Urteil der Fachgemeinschaft zugänglich, die gewonnenen Resultate werden überprüfbar.

Originalmitteilungen sollen die Gewinnung neuen Wissens kommunizieren und dokumentieren. Autoren müssen also nachweisen, was an ihrer Forschung neu ist. Eines muss neu sein, Problem oder Methode – oder zumindest die Kombination beider. Dieser Nachweis ist ohne Bezug auf früher publizierte Resultate nicht möglich. Frühere Arbeiten müssen auch zitiert werden, wenn Autoren in ihnen mitgeteiltes neues Wissen benutzt haben – ansonsten würden sie sich dem Vorwurf des Plagiats aussetzen oder wenigstens dem der Unkenntnis der Literatur. Dinge, die bereits in Lehrbüchern stehen, brauchen nicht original referenziert zu werden.

Neben der Neuheit ist auch die Relevanz des Problems und die Angemessenheit der Methoden nachzuweisen. Dazu kann man sich auf bereits publizierte Erfahrungen und Einschätzungen berufen. So enthalten viele Aufsätze einen Abriss der Entwicklung des Problem-Methoden-Gefüges, um das es den Autoren geht. Dies bietet zugleich die Möglichkeit, das Forschungsgebiet mit zu formen, indem durch das Setzen von Schwerpunkten die Richtung der weiteren Forschung beeinflusst wird. Den Lesern wird dadurch das Verständnis der Resultate erleichtert. Das gilt besonders für diejenigen, die in dem Gebiet nicht zu hause sind. Überblicksartikel (*reviews*) beschränken sich auf eine wertende und einordnende Darstellung bereits veröffentlichter Forschungsarbeiten und dokumentieren keine neuen Resultate.

1.2 Metadaten und Volltexte

Die in einer Bibliographie aufgeführten Dokumente müssen durch Angabe bestimmter Daten so charakterisiert werden, dass sie über das System der Literaturversorgung – auf das heute weitgehend über das World Wide Web zugegriffen werden kann – gefunden werden können. Meist werden dazu Erstautoren, die Erscheinungsjahre und bei Zeitschriftenaufsätzen die Namen der Zeitschriften, sowie Bandangaben (*volume*) und Anfangsseiten (*pages*) aufgelistet.

In bibliographischen Datenbanken werden außer diesem minimalen Satz von Daten noch solche aufgeführt und recherchierbar gemacht, die sich auf den Inhalt des jeweiligen Dokuments beziehen. Dazu gehören bei Zeitschriftenaufsätzen Titel, Schlüsselwörter (*keywords*, auch Schlagwörter oder Deskriptoren genannt) und Zusammenfassungen (*abstracts*). Die Institutionen der Autoren von Artikeln können ebenfalls für die Recherche von Nutzen sein.

In den 1960er Jahren entstand mit dem *Science Citation Index* (SCI) die erste bibliographische Datenbank, die auch die in den Zeitschriftenaufsätzen zitierten Quellen erfasst. Sein Schöpfer, Eugene Garfield, wollte die in den Referenzenlisten enthaltene Expertise der Autoren für das *Information Retrieval* nutzen. Er sah aber auch gleich, welche neuen Möglichkeiten für die quantitative Analyse der Wissenschaftsentwicklung durch den SCI gegeben waren. Er ist immer noch die am häufigsten für bibliometrische Untersuchungen genutzte Datenbank. Ich vermute aber, dass der Anteil von Aufsätzen, die auf seinen Daten basieren, messbar zurückgeht, sind ihm doch inzwischen nicht nur Schwesterdatenbanken in den Sozial- und Geisteswissenschaften erwachsen, sondern neuerdings auch vergleichbare Konkurrenten auf internationaler wie nationaler Ebene.

Heute kann über die Datenbankoberfläche außer auf Angaben *über* das Dokument (und über die in ihm zitierten Quellen), die man *Metadaten* nennt, auch schon oft auf das elektronische Do-

kument selber zugegriffen werden, und was noch wesentlicher ist, in seinem Volltext recherchiert werden. Der Datenbankrecherche noch nicht direkt zugänglich sind Bestandteile von Dokumenten, die die beiden Dimensionen des Lesemediums ausgeprägter beanspruchen als der sequentielle Text, nämlich Abbildungen und Tabellen. Deren Metadaten (Bildunter- und Tabellenüberschriften) dienen dann ihrem Auffinden. Die Auflösung aller im Dokument enthaltenen Information in eine Folge von Bits ermöglicht aber nicht nur prinzipiell auch eine (technisch schon realisierte) Suche in Bildern, sondern auch die Hinzunahme der dritten Raumdimension, der Zeitdimension oder beider. Digitale räumliche Modelle von Menschen und Gebäuden, sowie Bildsequenzen können in einigen Fällen nicht nur eine effiziente Form der Speicherung wissenschaftlichen Wissens darstellen, sondern auch eine adäquatere Visualisierung von Objekten und Prozessen ermöglichen.

1.3 Historisches

Bibliometrie – so wie der Begriff 1969 vorgeschlagen wurde und noch heute verwendet wird – hat nicht alle quantitativ erfassbaren Aspekte der Wissenschaftsentwicklung zum Gegenstand, sondern nur die den Output der Wissensproduktion betreffenden.[2] Mit dem Input, nämlich den für die Forschung benötigten Menschen, Geräten, Gebäuden und der Infrastruktur, beschäftigt sich traditionell die Wissenschaftsstatistik. Scientometrie – ebenfalls 1969 als Bezeichung vorgeschlagen (Nalimov und Mul'čenko 1969) – umfasst beides, Input wie Output, und hat darüber hinaus auch die wissenschaftliche Ausbildung im Blick.[3]

Bibliometrie geht über Scientometrie hinaus, insofern sie auch Literatur außerhalb der Wissenschaften untersucht, wenn auch selten: beide Gebiete überlappen sich weitgehend.

Der Begriff Bibliometrie wurde in den 1970-er Jahren zu dem der Informetrie verallgemeinert, welche als das Teilgebiet der Bibliotheks- und Informationswissenschaft angesehen werden kann, das alle quantitativen Seiten von Kommunikationsprozessen behandelt. Ein bibliometrisch inspiriertes Forschungsgebiet ist die Webometrie, die ebenfalls Teil der Informetrie ist. Insoweit sie wissenschaftliche Areale des Web untersucht, trägt sie auch zur Scientometrie bei. Sie hat dabei mit weitaus mehr Schwierigkeiten zu kämpfen als die Bibliometrie, weil aus dem Web gewinnbare Daten zur Wissenschaftskommunikation uneinheitlicher sind als die in bibliographischen Datenbanken erfassten.

Lange bevor der Begriff Bibliometrie geprägt wurde, publizierte der in den USA lebende Mathematiker Alfred Lotka (1926) seine für das Gebiet paradigmatische Untersuchung zur Produktivität von Autoren gemessen an der Zahl von Einträgen in zwei Fachbibliographien, einer physikalischen und einer chemischen (vgl. Abschnitt 2.1, S. 13). Er fand wenige Autoren mit vielen Einträgen und viele, die nur ein- oder zweimal vorkamen. Die Verteilung der Publikationen auf Autoren ist extrem schief und folgt weitgehend einer Potenzfunktion.

Zu einem ganz ähnlichen Ergebnis kam wenige Jahre später der Londoner Bibliothekar Samuel Bradford (1934) bei der Untersuchung der Verteilungen von Artikeln auf Zeitschriften in den Bibliographien zweier Spezialgebiete: wenige Kernzeitschriften enthalten den größten Teil der Literatur, während eine ganze Reihe von Zeitschriften im betrachteten Zeitraum jeweils nur einen Aufsatz zum Thema herausbringt. Das ist der wesentliche Inhalt des nach ihm benannten Gesetzes der Streuung von Literatur (*Bradford's law of scattering*, vgl. Abschnitt 2.3, S. 16).

Diese beiden wichtigen Arbeiten gehören zur Vorgeschichte der Bibliometrie. Damit sie ein Forschungsgebiet mit eigenem Namen werden konnte, musste erst der SCI geschaffen werden und Derek de Solla Price (1963) sein berühmtes Buch *Little Science, Big Science* verfassen. Er propagierte in ihm, die empirische Methodologie der Naturwissenschaften bei der Analyse ihrer eigenen Entwicklung anzuwenden und gab Beispiele für dieses Herangehen. Am bekanntesten ist das von ihm damals festgestellte exponentielle Wachstum der neuzeitlichen Wissenschaft (vgl. Abschnitt 2.6, S. 21).

De Solla Price war es auch, der schon früh erkannte, dass durch die Erfassung der Zitierungsbeziehungen zwischen Zeitschriftenaufsätzen in Datenbanken *Networks of Scientific Papers* (so der Titel der Arbeit in der Zeitschrift *Science* von 1965) der Analyse zugänglich werden (vgl. Abschnitt 3.1, S. 27). Die erste Zitationsanalyse wurde – wesentlich früher – von Gross und Gross (1927) publiziert.

Seit 1979 wurden in der damals gegründeten Zeitschrift *Scientometrics* vor allem bibliometrische Untersuchungen zur Wissenschaftsentwick-

[2] 'the application of mathematical and statistical methods to books and other media of communication' (Pritchard 1969)

[3] Nachzulesen in der Beschreibung des Fachgebietes auf der Titelseite der 1979 gegründeten Zeitschrift *Scientometrics: 'An International Journal for all Quantitative Aspects of the Science of Science, Communication in Science and Science Policy'.*

lung publiziert, so dass die Bezeichnungen Scientometrie und Bibliometrie oft synonym verwendet werden. 1987 begann in Diepenbeek (Belgien) unter dem Titel *Bibliometrics and Theoretical Aspects of Information Retrieval* eine Serie von internationalen Konferenzen, die seit der Gründung der *International Society for Scientometrics and Informetrics* (ISSI) 1993 in Berlin von ihr in zweijährigem Rhythmus organisiert wird.[4]

1.4 Methodisches

Die in Zeitschriftenaufsätzen publizierten Forschungsresultate tragen in ganz unterschiedlichem Maße zum Wissenschaftsfortschritt bei. Es ist deshalb zu fragen, wie Publikationsstatistiken überhaupt sinnvoll interpretiert werden können. Als wie verschieden sich auch die Bedeutung eines Forschungsresultats herausstellt – sie steht keinesweg von Anfang an und für immer fest – so zeigt die Publikation doch an, dass Verfasser, Gutachter und Herausgeber ihm ein notwendiges Minimum an Bedeutung zugemessen haben. Davon kann bei der Interpretation statistischer Verteilungen von Publikationszahlen ausgegangen werden, ganz so wie bei der Interpretation von Bevölkerungsstatistiken, bei denen auch höchst unterschiedliche Individuen alle als Einwohner einer Stadt, z. B., gezählt werden. Und wie eine Bevölkerungstruktur nach Geschlecht, Alter, Einkommen usw. erhoben werden kann, so haben auch Journalaufsätze Merkmale, die es lohnt, in die Analyse einzubeziehen. Was die Bedeutsamkeit für den Wissenschaftsfortschritt angeht, wäre da vor allem die Rezeption der Publikation zu nennen, wie sie sich in der Zitationsgeschichte widerspiegelt, welche als zeitliche, fachliche, geographische usw. Verteilung der Zitierungszahlen quantitativ fassbar wird.

Grundlage aller weiteren Aggregation und Interpretation ist zuerst die Ermittlung von Häufigkeitsverteilungen durch einfaches Zählen von Publikationen, Zitierungen, Autoren usw. Die Verteilungen können dann durch geeignete statistische Maßzahlen, wie Mittelwert, Standardabweichung und Schiefe charakterisiert werden.

Die nächste Frage ist dann, ob sich die empirischen Verteilungen durch mathematische Funktionen beschreiben lassen. So gewinnt man theoretische Verteilungen, deren Parameter an die empirischen Daten angepasst werden und die als phänomenologische Modelle des untersuchten Ausschnitts der Wissenschaftskommunikation angesehen werden können.

Über die nur beschreibenden phänomenologischen Modelle gehen solche hinaus, die konstruiert werden, um die empirischen Befunde zu erklären, d. h. sie auf einfache, allgemein einsehbare Annahmen zurückzuführen. Erklärende bibliometrische Modelle sind Bausteine einer noch zu schaffenden empirisch fundierten Theorie der Wissenschaft. Von Wissenschaftsforschern und Wissenschaftssoziologen entwickelte theoretische Ansätze müssen sich auch an bibliometrischen Forschungsresultaten bewähren, d. h. ihnen erstens nicht widersprechen und darüber hinaus zu ihrem Verstehen beitragen.

Aus der Theorie heraus interessierende und durch sie zu definierende Begriffe, wie z. B. die Produktivität von Forschern oder die Forschungsvielfalt in einem Fachgebiet, müssen in bibliometrische Indikatoren übersetzt werden, d. h. in auf bestimmte Weise zu gewinnende mehr oder minder komplexe statistische Maßzahlen.

Dieser Teil bibliometrischer Forschung ist besonders für die wissenschaftspolitische Anwendung wichtig. Es geht um möglichst einfache aber aussagekräftige bibliometrische Indikatoren der Wissenschaftsentwicklung, analog zur Wirtschaft, wo für die Wirtschaftspolitik ökonomische Indikatoren bereitgestellt werden.

Neben Verteilungen, Modellen und Indikatoren sind in den letzten Jahren Begriffe aus der Analyse sozialer Netzwerke (*Social Network Analysis*, SNA) für die Bibliometrie immer interessanter geworden, weil die schon lange diskutierten bibliometrisch erfassbaren Netzwerke von Artikeln, Autoren, Instituten, z. B., mittlerweile rechentechnisch besser zu bewältigen sind. Das Web analysierende Informatiker und statistische Physiker konnten teilweise auch auf bibliometrische Begriffsbildungen zurückgreifen. Weil bibliometrische Netzwerke und das Web so groß sind, müssen für ihre Analyse die auf kleine soziale Netzwerke zielenden SNA-Methoden um statistische ergänzt werden.

1.5 Literatur

Das erste einschlägige Lehrbuch wurde von Egghe und Rousseau (1990) unter dem Titel *Introduction to Informetrics* veröffentlicht. Aus diesem Buch habe ich viel gelernt, insbesondere der mathematischen Seite der Bibliometrie kann man durch seine Lektüre näher

[4] `http://www.issi-society.info`

kommen.[5] Kürzlich hat Leo Egghe (2005) eine ausgesprochen mathematische Monographie zu *Power laws in the information production process* geschrieben. Sein Ziel war es, alle aus der Annahme einer Potenzfunktion für die Häufigkeitsverteilung ableitbaren und für die Informetrie wichtigen mathematischen Beziehungen zusammenzustellen. Auch das online frei verfügbare Vorlesungsskript von Wolfgang Glänzel (2003) zum Thema *Bibliometrics as a Research Field* ist dadurch geprägt, dass sein Autor Mathematiker ist.

Ein wichtiger aktueller Sammelband zur Nutzung von Publikations- und Patentstatistiken für die Wissenschafts- und Technikforschung wurde von Moed, Glänzel und Schmoch (2004) herausgegeben. Ein brauchbares deutschsprachiges Lehrbuch lag bisher nicht vor.

[5] Es ist seit einiger Zeit als elektronisches Buch online frei verfügbar.

Kapitel 2

Bibliometrische Verteilungen

2.1 Die Lotka-Verteilung

1926 veröffentlichte Alfred Lotka, ein namhafter Biomathematiker und Statistiker, zu dieser Zeit bei einer New Yorker Versicherung beschäftigt, seinen berühmten Artikel zur Produktivität wissenschaftlicher Autoren. Er leitet ihn mit folgendem Satz ein: *It would be of interest to determine, if possible, the part which men of different calibre contribute to the progress of science.*

Es ist sicher kein Zufall, dass ein Statistiker, der eine solche Frage aufwirft, bei einer Versicherungsgesellschaft tätig ist, sind Versicherungen doch daran interessiert, soziales Verhalten quantitativ zu beschreiben und mögliche Regelhaftigkeiten darin zu erkennen, um Versicherungsrisiken besser abschätzen zu können. Dabei sind oft Häufigkeitsverteilungen von Nutzen. Als einfachstes Beispiel kann die Altersverteilung der Bevölkerung dienen.

Lotka interessierte sich nun für die Häufigkeitsverteilung wissenschaftlicher Produktivität, so der Titel seines Aufsatzes: *The frequency distribution of scientific productivity.* Er wollte also wissen, wie häufig Wissenschaftler einer bestimmten Produktivität anzutreffen sind. Die Produkte von Wissenschaftlern sind vor allem Publikationen. Diese können ganz unterschiedlich zum Wissenschaftsfortschritt beitragen, als erstes, grobes Maß für die Produktivität eines Autors ist die Anzahl seiner Publikationen während einer Zeitspanne jedoch brauchbar (vgl. Abschnitt 1.4, S. 10).

Lotka fand erstaunlicherweise mathematisch einfach beschreibbare und ganz ähnliche Verteilungen in zwei ganz unterschiedlichen Fachbibliographien, und zwar in den *Chemical Abstracts* von 1907 bis 1916 und in *Auerbachs Geschichtstafeln der Physik* (Leipzig 1910). Letztere enthalten nur herausragende Beiträge, die tatsächlich zum Wissenschaftsfortschritt wesentlich beigetragen haben.

In beiden Bibliographien sind die meisten Autoren (ca. 60%) nur mit einem Beitrag vertreten. Zu höheren Publikationszahlen hin nimmt die Zahl der Autoren rapide ab: Viele Autoren publizieren wenig, nur wenige Autoren viel (s. Abb. 2.1). Eine solche schiefe Verteilung ist auch bei Vermögen bekannt: Wenige besitzen viel, viele nur wenig. Diese Vermögensverteilung war zu Lotkas Zeiten bereits von Ökonomen quantitativ untersucht und näherungsweise durch eine Potenzfunktion beschrieben worden, wie man in Lotkas Artikel nachlesen kann (unten mehr dazu). Genau dies versucht nun Lotka auch im Falle der Publikationszahlen von Autoren. Wenn man die Häufigkeit (*frequency*) von Autoren mit j Publikationen als $f(j)$ bezeichnet, dann lässt sich eine Potenzfunktion mit fallender Tendenz schreiben als

$$f(j) = \frac{C}{j^\alpha}, \tag{2.1}$$

wobei der Exponent $\alpha > 0$ und die Konstante C aus den empirischen Daten zu bestimmende Parameter sind. Dies soll vernünftigerweise so erfolgen, dass die theoretischen Erwartungswerte für die Häufigkeiten den empirischen Werten möglichst nahe kommen.

Potenzfunktionen haben nun die schöne Eigenschaft, dass sie in doppelt logarithmischer Darstellung als gerade Linie erscheinen.[1] So kann man schnell nachprüfen, ob die empirischen Daten einem Potenzgesetz (*power law*) folgen. Lotka fand, dass die Zahlen von Autoren mit 1, 2, 3, ... Publikationen für beide Bibliographien nahe einer Geraden mit $\alpha = 2$ liegen (s. Abb. 2.1).

Durch eine Reihe von Punkten möglichst optimal eine Gerade zu legen, ist eine relativ einfache mathematische Aufgabe, die auch ohne Rechenmaschine zu erledigen ist. Die dafür geeigneten Formeln für die Parameter α und C lassen sich ableiten, wenn man die Differenzen zwischen beobachteten und theoretischen Werten quadriert

[1] Weil $\log f(j) = \log C - \alpha \log j$ eine Gleichung für eine Gerade darstellt. Ihr Anstieg ist $-\alpha$, bei $\log C$ schneidet sie die vertikale Achse.

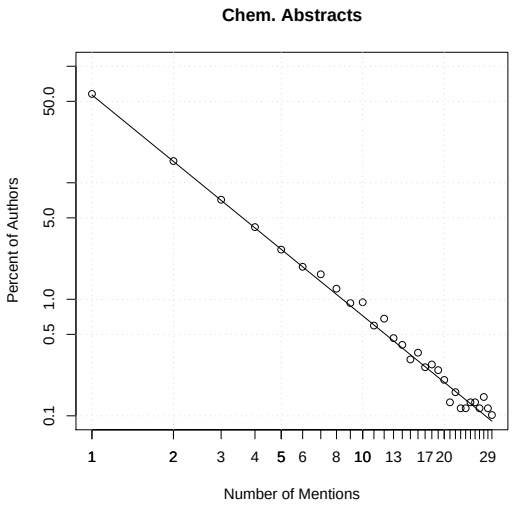


Abbildung 2.1: Das Lotka-Diagramm der *Chemical Abstracts* 1907–1916 (Buchstaben A und B) in doppelt-logarithmischer Darstellung. Die Gerade entspricht einer Potenzfunktion mit $\alpha = 1.888$. Quelle: Lotka (1926)

und fordert, dass die Summe der Quadrate der Differenzen minimal wird.[2]

Wie gut die empirischen Resultate durch die Gerade beschrieben werden, lässt sich am Korrelationskoeffizienten R ablesen ($-1 \leq R \leq 1$). Sein Quadrat gibt an, welchen Anteil der Variation der empirischen Daten in vertikaler Richtung durch die Gerade beschrieben wird. Ist R^2 klein, so lassen sich auch die Parameter α und C nur ungenau bestimmen, was durch Angabe ihrer Fehlergrenzen deutlich gemacht werden kann. Lotka fand mittels linearer Regression die Werte 2.021 ± 0.017 für die Auerbach'schen Geschichtstafeln und 1.888 ± 0.007 für die *Chemical Abstracts*.[3]

Die Bedeutung von Lotkas Arbeit besteht darin, ein erstes Beispiel statistischer Analyse von wissenschaftlichen Bibliographien gegeben zu haben, welche erstaunlicherweise eine einfache mathematische Beschreibung mittels einer Potenzfunktion ergab. Untersuchungen weiterer Bibliographien haben immer wieder Lotkas Ergebnisse bestätigt, wenn auch der Parameter α variiert. Erstaunlich ist auch – was schon Lotka selbst feststellte – dass die Auerbach'schen Geschichtstafeln (mit nur herausragenden Leistungen in der Physik über Jahrhunderte) dem gleichen Gesetz folgen, wie die relativ vollständige Bibliographie der Chemie über zehn Jahre. Erklären konnte Lotka das von ihm aufgedeckte Phänomen nicht, Erklärungsversuche blieben aber nicht aus. Sie werden im Kapitel über bibliometrische Modelle abgehandelt.

Drei Anmerkungen müssen noch gemacht werden. Die erste betrifft die Behandlung von Beiträgen mit mehreren Autoren. Lotka hat bei ihnen jeweils nur den ersten Autor berücksichtigt (s. Fußnote 5 von Lotkas Arbeit). Im 20. Jahrhundert hat aber in allen Wissenschaftsgebieten die Koautorschaft so stark zugenommen, dass dieses vereinfachende Verfahren zur Bestimmung der Produktivität von Autoren nicht mehr angemessen ist. Dazu wird im Kapitel über bibliometrische Indikatoren Näheres ausgeführt (vgl. auch Abschnitt 2.4, S. 18).

Die zweite Anmerkung betrifft die Methode der Parameterschätzung. Die Anpassung einer Geraden an die logarithmierten Werte ergibt andere Parameterwerte als eine direkte Anpassung der Potenzfunktion an die nicht logarithmierten Originaldaten, wie man sie heute leicht iterativ mit einem Rechner durchführen kann. Trotzdem wird mit einiger Berechtigung auch heute noch die lineare Regression log Werte für die Parameterberechnung von Potenzfunktionen verwendet. Die Berechtigung liegt darin, dass es für Autoren ähnlich schwierig ist, ihre Publikationszahl (in einer zeitlich abgeschlossenen Bibliographie) von fünf auf sechs wie von 50 auf 60 Beiträge zu steigern. Die multiplikative Steigerungsrate ist in beiden Fällen 6/5, was für die logarithmierten Werte einen additiven Zuwachs von $\log 5/6$ bedeutet. Man kann sagen, dass der Logarithmus der Publikationszahl ein adäquaterer Indikator der Produktivität von Autoren ist als die Publikationszahl selbst.[4]

Noch eine Bemerkung zu in anderen Fachgebieten gefundenen schiefen GrößenVerbTeigLungen, die einem Potenzgesetz folgen. Die Vermögensverteilung wurde oben schon erwähnt. Sie gleicht oft einer Pareto-Verteilung,[5]

[2] Diese so genannte *lineare Regression* entspricht einem physikalischen Modell, bei dem zwischen den empirischen Punkten im Diagramm und der Geraden in vertikaler Richtung Zugfedern angebracht sind, welche die Gerade optimal positionieren.

[3] Lotka berücksichtigte bei den Berechnungen nicht alle empirischen Daten. Die am meisten produktiven Autoren gingen nicht in seine Rechnung ein, sondern nur die Autoren bis höchstens 17 Publikationen für die Daten von Auerbach und die bis höchstens 30 für die Chemie: *Beyond this point fluctuations become excessive owing to limited number of persons in the sample.* (s. Fußnote 1 von Lotkas Artikel).

Ich benutze in diesem Text den Punkt als Dezimalzeichen.

[4] Für die auf der vertikalen Achse aufgetragene Zahl von Autoren mit 1, 2, ... Publikationen kann analog argumentiert werden.

[5] Vilfredo Federico Pareto (1848–1923) war ein italienischer Ökonom und Soziologe (Quelle: Wikipedia).

die mathematisch durch eine unten abgeschnittene Potenzfunktion definiert ist.

In der Linguistik kennt man das Zipf'sche Gesetz der Worthäufigkeiten, ebenfalls ein Potenzgesetz.

Zusammenfassung: Die Lotka-Verteilung wissenschaftlicher Autoren nach ihrer Produktivität folgt einem Potenzgesetz, nach dem die Zahl $f(j)$ der Autoren mit j Beiträgen zu einer Fachbibliographie mit $1/j^{\alpha}$ abnimmt. Lotka fand $\alpha \approx 2$ und für den Anteil von Autoren mit nur einem Beitrag Werte nahe 60%.

2.2 Exkurs: Gesetz von Zipf

Der US-amerikanische Linguist George Kingsley Zipf (1902–1950) führte in den dreißiger Jahren quantitative Methoden in sein Fachgebiet ein. Er untersuchte z. B., ob häufig gebrauchte Wörter durchschnittlich weniger Silben oder Phoneme aufweisen (Zipf 1935).[6] Er ermittelte tatsächlich für mehrere Sprachen einen solchen Zusammenhang und begründete damit seine These, dass in der sprachlichen Evolution ein Prinzip des kleinsten Aufwands (*principle of least effort*) wirke (Zipf 1949). Er stützte sich dabei auf Kompilationen von Worthäufigkeiten, die er auch – ganz analog zu Lotka – in Bezug auf ihre Verteilung untersuchte. Er erhielt auch das gleiche Ergebnis wie Lotka: Wenige Worte werden viel gebraucht, viele wenig und die Verteilung der Wörter nach ihrer Häufigkeit folgt einem Potenzgesetz mit dem Exponenten $\alpha \approx 2$ (vgl. Formel 2.1, S. 13).

Als Beispiel soll die Verteilung der Wörter in der Umgangssprache der Bewohner von Beijing (das damals Peiping hieß) dienen (Zipf 1932). Die Verteilung der am wenigsten häufigen Wörter ist zusammen mit der durch lineare Regression der logarithmierten Werte ermittelten Regressionsgeraden in Abbildung 2.2 zu sehen. Wie Lotka hat auch Zipf hier die Regression auf kleine Häufigkeiten beschränkt. Er ging aber darüber hinaus und hat in eine veränderte Darstellung auch die sehr häufigen Wörter einbezogen. In den heute als *Zipf-Plot* bezeichneten Diagrammen wird statt der Zahl der Wörter ihr jeweiliger Rang aufgetragen.[7] In Abbildung 2.3 ist das häufigste Wort rechts unten zu finden (Rang 1 auf der y-Achse, Häufigkeit: 905). Leider hat Zipf die Häufigkeiten der nächsten elf

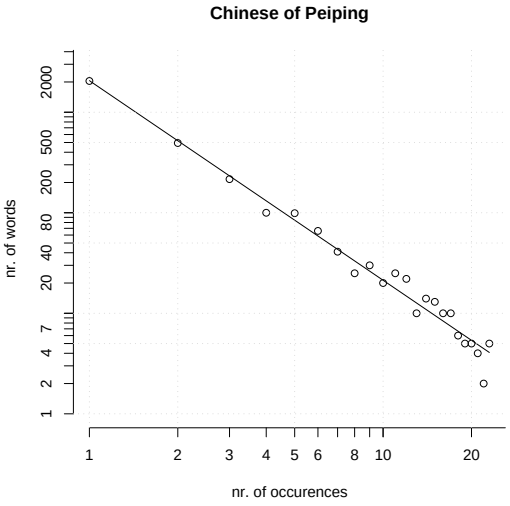


Abbildung 2.2: Doppelt logarithmische Darstellung der Verteilung der am wenigsten häufigen Wörter (bis zu maximal 23-maligem Auftreten in zwanzig Stichproben verbundener Umgangssprache zu je tausend Silben). Exponent der Potenzfunktion $\alpha = 1.985 \pm .071$, $R^2 = 0.97$. Quelle der Rohdaten: Zipf (1935)

Ränge nicht in seine Tabelle aufgenommen (Zipf 1935, S. 26). Deswegen ist als nächstes erst der dreizehnte Rang mit Häufigkeit 101 dargestellt. Als erstaunliches Ergebnis erhalten wir, dass

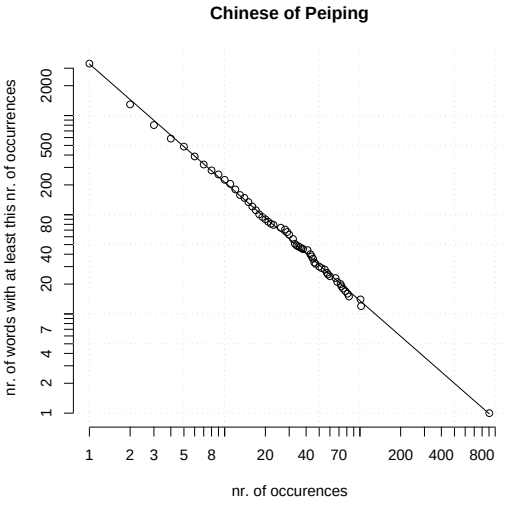


Abbildung 2.3: Zipf-Plot: Doppelt logarithmische Darstellung der Verteilung aller Wörter (Stichproben wie oben). Exponent der Potenzfunktion $\gamma = 1.193 \pm .007$, $R^2 = 0.998$. Quelle der Rohdaten: Zipf (1935)

[6] vgl. den Aufsatz von Ronald Rousseau (2002)

[7] "As suggested by a friend, one can consider the words of a vocabulary as ranked in the order of their frequency, e.g. the most frequent word, the second most frequent, the third most frequent, the five-hundreth most frequent, the thousenth most frequent, etc." (Zipf 1935, S. 44–45)

auch die Rang-Häufigkeitsverteilung im Zipf-Plot annähernd einem Potenzgesetz folgt, allerdings mit einem kleineren Exponenten als die Häufigkeitsverteilung selber.

Dieses Ergebnis kann man sich folgendermaßen erklären. Die y-Achse in der Abbildung 2.3 hat streng genommen eigentlich nur für die ersten ca. 50 Ränge die Bedeutung eines Rangs (da wo nur ein bis zwei Wörter auf einem Rang sitzen). Am anderen Ende der nur einmal, zweimal etc. auftretenden Wörter gibt es nur einen Punkt für 2046, 494 etc. Wörter, nämlich den mit dem jeweils letzten Rang dieser Wörtergruppe. Eine andere Interpretation des Diagramms ist daher viel einleuchtender und auch als Achsenbeschriftung angezeigt: Es handelt sich bei der y-Achse im Zipf-Plot um die Zahl der Wörter, die mindestens so oft auftreten, wie auf der x-Achse angegeben. Man trägt nicht die Häufigkeiten $f(j)$ auf, sondern

$$F(i) = \sum_{j \geq i} f(j) = \sum_{j=i}^{\max(i)} f(j). \tag{2.2}$$

Für relative Häufigkeiten $p(j) = f(j)/T$ ($T = \sum_{j=1}^{\max(j)} f(j)$) wird das auch als *kumulative Verteilungsfunktion* bezeichnet:

$$P(i) = \sum_{j=i}^{\max(i)} p(j). \tag{2.3}$$

Gerade für große Häufigkeiten i kann man in guter Näherung diese Gleichung auch mit kontinuierlich aufgefasster Variable j schreiben, wodurch sich die Summe in ein Integral verwandelt:

$$P(i) = \int_{i}^{\max(i)} p(j)\mathrm{d}j. \tag{2.4}$$

Für eine Potenzfunktion der Häufigkeitsverteilung $p(j)$ ergibt sich für die kumulative Verteilungsfunktion $P(i)$ nur dann ein Potenzgesetz, wenn $\max(i) = \infty$ gesetzt werden kann, was unrealistisch ist. Aus der Lotka-Verteilung folgt also i. a. nicht das Zipf-Gesetz. Umgekehrt folgt in der kontinuierlichen Näherung aus 'Zipf' jedoch 'Lotka' (Rousseau 2002, S. 15–16). Um das einzusehen, differenzieren wir die vorige Gleichung nach i und benutzen dabei die Regel, dass die Ableitung eines Integrals nach der unteren Grenze gleich dem negativen Argument des Integrals ist: $P'(i) = -p(i)$. Andererseits gilt für die Ableitung der Potenzfunktion $P(i) = c/i^{\gamma}$ nach i: $P'(i) = -\gamma c/i^{\gamma+1}$. Also folgt aus dem Zipf-Gesetz eine Lotka-Verteilung mit einem um 1 erhöhten Exponenten ($\gamma + 1 = \alpha$, $\gamma c = C$):

$$p(i) = \frac{\gamma c}{i^{\gamma+1}} = \frac{C}{i^{\alpha}}. \tag{2.5}$$

Diese Beziehung zwischen den Exponenten γ und α ist in unserem Beispiel vor allem wegen der Näherung (aber sicher auch wegen der jeweils eingeschränkten Datenbasis) nicht exakt erfüllt ($\gamma \approx 1.2$, $\alpha \approx 2$).

Die Analyse von Worthäufigkeiten ist nicht im eigentlichen und traditionellen Sinne Gegenstand der Bibliometrie.[8] Dieses Kapitel ist also ein erster Exkurs über ein informetrisches Thema, dem weitere folgen werden.

2.3 Die Bradford-Verteilung

Samuel C. Bradford (1878–1948), von 1925 bis 1937 Direktor der *Science Museum Library* in London,[9] schätzte in seinem Aufsatz von 1934 ab,[10] dass damals die 300 spezialisierten Referate-Zeitschriften (*abstracting and indexing journals*) für wissenschaftliche und technische Literatur nur rund ein Drittel der relevanten Publikationen erfassten. Er vermutete die Ursache dieses Mangels in der breiten Streuung der relevanten Literatur über eine große Anzahl von Quellen und bestätigte diese Vermutung anhand von zwei Bibliographien, die von einem Mitarbeiter erstellt worden waren.

Ein Kern von einigen Zeitschriften enthält einen großen Teil der gesuchten Literatur und eine große Zahl von Journalen enthält nur ab und zu einmal einen Artikel zum Thema. Das ist es, was den Aufwand so hoch treibt, will man eine vollständige Bibliographie zu einem Thema zusammenstellen.

Bradfords praktische Schlussfolgerung war, die spezialisierten Referate-Dienste abzuschaffen zugunsten eines Systems, das alle wissenschaftlichen und technischen Periodika indexiert und die Nachweise dann nach Fachgebieten sortiert.[11]

Um sein Zahlenmaterial quantitativ auszuwerten, ordnete Bradford die Journale in eine Rangfolge nach der Zahl der Artikel. Das später *Bradford's law of scattering* genannte Gesetz der Literaturstreuung formulierte er so: Wenn die wissenschaftlichen Journale in der Rangfolge fallender Produktivität von Artikeln in einem Fachgebiet geordnet werden, dann können sie unterteilt werden in einen Kern (*nucleus*) von

[8] Wenn man nicht die Untersuchung von Volltexten zu ihrem Gegenstandsbereich zählt (vgl. Abschnitt 1.2, S. 8).

[9] Quelle: Wikipedia

[10] Der Artikel wurde 1985 nachgedruckt, s. Bibliographie.

[11] Die Idee eines fachübergreifenden Indexes wurde von Garfield bei der Konzipierung des *Science Citation Index* aufgegriffen (vgl. Abschnitt 1.2, S. 8).

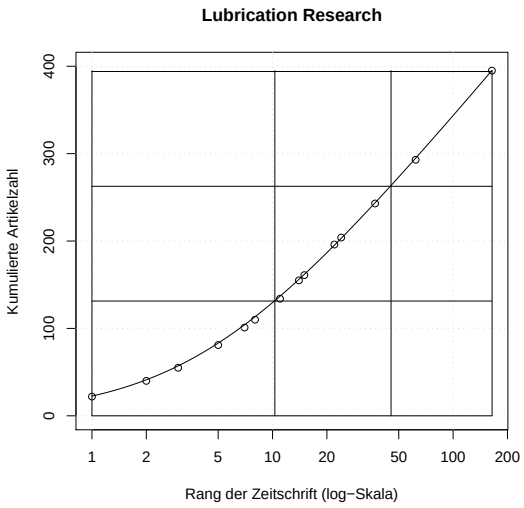


Abbildung 2.4: Das Bradford-Diagramm der Schmiermittel-Forschung 1931–1933 (halb-logarithmische Darstellung). Die Kurve entspricht einer Leimkuhler-Funktion mit $a = 107.7$ und $b = 0.233$. Die waagerechten Linien unterteilen die Artikel in drei gleich große Teilmengen, die senkrechten zeigen den Kern und zwei Zonen von Zeitschriften (s. Text).

N_0 Periodika, die sich speziell dem Forschungsgegenstand widmen und mehreren Zonen, die die gleiche Zahl von Artikeln umfassen wie der Kern, während die Zahl der Periodika von Zone zu Zone stark anwächst, und zwar nach dem Gesetz $N_0, N_0k, N_0k^2, \ldots$:

$$N_n = N_0k^n. \tag{2.6}$$

Wir veranschaulichen uns dieses Gesetz anhand von Bradfords eigenen Daten, die er aus der Bibliographie zum Thema Schmiermittel (*lubrication*) gewonnen hat. Die Verteilung ist in Abbildung 2.4 dargestellt. Die horizontale Achse des Diagramms verzeichnet in logarithmischer Skalierung den Rang des jeweiligen Journals bezüglich seiner Artikelzahl in der Bibliographie. Die vertikale Achse gibt die kumulative Zahl der Artikel in den Journalen bis zum jeweiligen Rang an. Zeitschriften mit gleicher Artikelzahl erhalten hier dennoch unterschiedliche Ränge zugeordnet.[12] Dadurch wird der Rang gleich der kumulativen Zahl von Zeitschriften. Bradford hat seine Ergebnisse – neben der verbalen Formulierung im Zonen-Modell – in genau einem solchen Diagramm dargestellt. Die waagerechten Linien teilen die Menge der 395 Artikel in drei gleiche Teile, die den drei Zonen von Zeitschriften entsprechen. Der Kern enthält zehn Zeitschriften, die beiden weiteren Zonen jeweils ungefähr das 3.5-fache der vorigen Zone ($N_0 = 10, k = 3.5, N_0k = 35, N_0k^2 = 122.5$).

Eine Funktion, die sich für die Beschreibung vieler empirischer Bradford-Graphen eignet, wurde von Leimkuhler (1967) vorgeschlagen:

$$R(r) = a \log(1 + br). \tag{2.7}$$

Die Leimkuhler-Funktion für $a = 107.7$ und $b = 0.233$ fittet Bradfords Schmiermittel-Daten dem Augenschein nach recht gut (s. Abb. 2.4).[13] Allgemein gilt, dass die Leimkuhler-Funktion sich für große Werte von r der Funktion $a \log(br) = a \log r + a \log b$ annähert. Dem entspricht in der Darstellung mit logarithmischer x-Achse eine Gerade mit dem Anstieg a, die bei $a \log b$ die y-Achse schneidet.[14] Weil Bradford die Annäherung seiner Datenpunkte an eine solche Gerade deutlich machen wollte, wählte er die halb-logarithmische Darstellung. Bei anderen Bibliographien ergibt sich aber oft ein Bradford-Diagramm, bei dem die Datenpunkte nicht wie hier nahe einer J-förmigen Kurve liegen, sondern eher bei einer $\int$-förmigen. Die Leimkuhler-Funktion kann dies nicht beschreiben. Sie wurde deshalb von Rousseau (1988) verallgemeinert.[15] Daneben wurde eine große Zahl alternativer Modelle vorgeschlagen.[16]

Bradfords oben erwähnte Argumentation zu den spezialisierten Referate-Diensten wird von ihm am Schluss seines Aufsatzes auch auf Spezialbibliotheken ausgedehnt: “special libraries cannot gather together the complete literature of their subject, except by relinquishing altogether their specific character and becoming practically general libraries of science” (Bradford 1934).

Bei begrenztem Platz und Budget für gedruckte Journale in einer Spezialbibliothek konnten Entscheidungen zum Bestellen von Fachzeitschriften leichter gefällt werden, wenn man diese – wie Bradford es tat – in eine Rangliste nach der Zahl ihrer Beiträge zum Spezialgebiet anordnete. Heute spielt Platz für Journale immer weniger eine Rolle, weil der Online-Zugriff auf Zeitschriftenaufsätze vorherrschend wird, und das Budget wird zum alleinigen Kriterium. Auch dieses wird unwichtig werden, sollten wissenschaftliche Zeitschrif-

[12] Als Punkt dargestellt wird jeweils immer nur die letzte Zeitschrift gleicher Artikelzahl (analog zum Zipf-Plot).

[13] bei Verwendung natürlicher Logarithmen

[14] Für $b < 1$ gilt $a \log b < 0$, d. h. die y-Achse wird im negativen Bereich von der asymptotischen Geraden geschnitten.

[15] s. Gleichung IV.5.2 (S. 341) im Lehrbuch von Egghe und Rousseau (1990)

[16] s. z. B. den Übersichtsartikel von Oluic-Vukovic (1997)

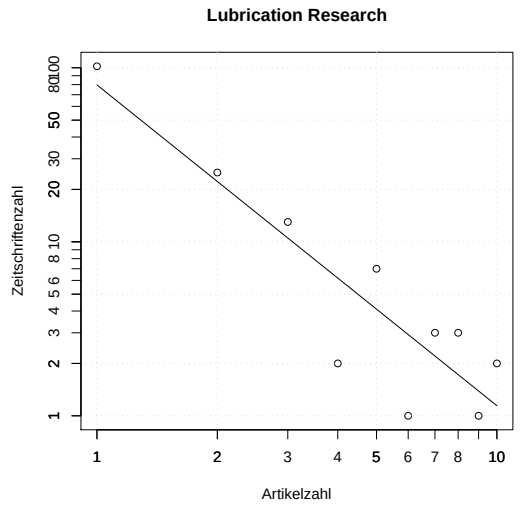


Abbildung 2.5: Verteilung der Zeitschriften nach der Zahl der Beiträge zur Schmiermittel-Forschung 1931–1933 (doppelt-logarithmische Darstellung). Die Gerade entspricht einer Potenzfunktion mit $\alpha = 1.84 \pm .31$ (s. Text).

ten mit freiem Zugang (*open access*) zum typischen Verbreitungsmedium für Forschungsergebnisse werden. Damit würde Bradfords Problem bezüglich der Spezialbibliotheken genauso gelöst werden, wie es für die Referate-Zeitschriften durch fachübergreifende bibliographische Datenbanken, wie den *Science Citation Index*, gelöst wurde.

Das *law of scattering* bleibt dennoch ein wichtiges Grundgesetz der Bibliotheks- und Informationswissenschaft – solange es Fachzeitschriften geben wird. Die exakte mathematische Modellierung stand für Bradford nicht im Mittelpunkt des Interesses; sie kann jedoch Aufschlüsse über spezifische Eigenschaften von Bibliographien in verschiedenen Fachgebieten liefern.

Das Bradford-Diagramm beginnt – im Gegensatz zum Lotka-Diagramm – mit den am meisten produktiven Quellen. Wenn Bradfords Daten zur Literaturstreuung in der Schmiermittelforschung so wie Lotkas Daten zur Produktivität von Autoren dargestellt werden, ergibt sich ein zur Lotka-Verteilung ganz ähnliches Bild. In Abb. 2.5 ist erkennbar, dass auch hier die Produktivität der Quellen angenähert nach einem Potenzgesetz verteilt ist (wenigstens die der am wenigsten produktiven Quellen). Sogar der (wieder durch lineare Regression der logarithmierten Werte gewonnene) Exponent des Gesetzes ist von ähnlicher Größe wie Lotkas, nur dass die Daten stärker streuen (ein Umstand, der im Bradford-Diagramm nicht zu Tage tritt, weil beide Achsen kumulative Skalen haben).

Zusammenfassung: Die Bradford-Verteilung wissenschaftlicher Fachzeitschriften nach der Zahl ihrer Beiträge zu einer Fachbibliographie lässt sich durch die Einteilung der Journale in einen Kern der am meisten zum Thema beitragenden und weitere Zonen von weniger beitragenden Periodika charakterisieren. Kern und Zonen können so gewählt werden, dass sie jeweils die gleiche Zahl von Artikeln enthalten, die aber über jeweils immer mehr Zeitschriften verstreut sind. *Bradford's law of scattering* ist in vielen Bibliographien so realisiert, dass sich die Zahlen der Zeitschriften in Kern und Zonen angenähert so verhalten wie $1 : k : k^2$... Das Problem einer allgemeingültigen mathematischen Modellierung empirischer Bradford-Verteilungen wird in der bibliometrischen Literatur noch diskutiert.

2.4 Lognormalverteilungen der Publikationsproduktivität

Misst man die Produktivität von wissenschaftlich Forschenden an der Zahl der Zeitschriftenartikel, bei denen sie als Autoren auftreten, dann kann man sich heute nicht mehr einfach nur auf die Erstautoren beschränken – wie es Lotka (1926) noch mit Berechtigung tat. Weil schon seit langem in vielen Forschungsgebieten Aufsätze mit nur einem Autor in der Minderheit sind, würde man so nur ein unvollständiges Bild erhalten. Eine Alternative zu Lotkas einfacher Zählweise (*straight counting*) besteht darin, für einen Artikel jedem seiner Autoren eine Publikation anzurechnen (*normal counting*). Zählt man nach dieser Methode, dann werden in großen Teams arbeitende Forscher mit vielen Koautoren tendenziell produktiver erscheinen als allein oder in einem kleinen Team arbeitende. Diese Verzerrung des Bildes wird durch die fraktionale Zählweise (*fractional counting*) vermieden. Bei ihr wird ein Artikel auf die Autoren aufgeteilt, so dass sie nur Anteile < 1 (*fractions*) angerechnet bekommen. Fraktionales Zählen verhindert, dass ein Artikel mit k Autoren k-mal in die Analyse eingeht. Der fraktionale Produktivitätsindikator f_i von Autor i ist dann die Summe seiner Artikelanteile. Da man die tatsächlichen Anteile von Autoren an einer gemeinsamen Publikation schlecht einschätzen kann, wird fast immer von einer Gleichverteilung ausgegangen, d. h. jeder der k Autoren erhält einen Wert von $1/k$ angerechnet.

Soll dieser Indikator für Forschungsgruppen oder Institute berechnet werden, genügt es, die Summe der f_i aller Mitarbeiter zu bilden. Der fraktionale Produktivitätsindikator ist additiv – im Gegensatz zu den normal gezählten Autorschaften.[17]

Die erste mir bekannte Untersuchung der Publikationsproduktivität, die auf fraktionaler Zählweise basiert, führte der US-amerikanische Physik-Nobelpreisträger William Shockley (1957) durch. Er fand, dass die fraktionale Produktivität von 88 Forschern am *Brookhaven National Laboratory* auf Long Island (USA) gemessen an Einträgen in den *Science Abstracts A & B*[18] für die Jahre 1950–1953 ziemlich genau einer *Lognormalverteilung* folgt.[19]

Eine Größe ist lognormal verteilt, wenn ihr Logarithmus einer Normalverteilung unterliegt. Empirische Daten sind oft normalverteilt. Als Beispiel kann hier die Größenverteilung von Jungen eines bestimmten Alters genannt werden. Deren Körpergröße streut symmetrisch (mit einer Varianz σ^2) um einen Mittelwert μ, wobei die Abnahme der Zahl von Jungen beiderseits des Mittelwerts durch die Gauß'sche Glockenkurve veranschaulicht wird, welche durch die *Dichtefunktion* der Normalverteilung mathematisch beschrieben wird:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \qquad (2.8)$$

Wenn der Logarithmus der Publikationsproduktivität normalverteilt ist, dann ergibt dies – wie bei Lotka und Bradford – eine schiefe Verteilung der Produktivität selber: wenige hochproduktive Quellen stehen vielen wenig produktiven gegenüber. Nach einem Potenzgesetz verteilte Größen – wie die Publikationszahlen bei Lotka – werden jedoch für kleine Werte immer häufiger, während die Dichtefunktion der Lognormalverteilung ein Maximum besitzt und bei der Annäherung an Null verschwindet (s. Abb. 2.6).

Die fraktionale Publikationsproduktivität f_i eines Autors i kann beliebige Werte ≥ 0 annehmen, sie ist eine quasi kontinuierliche Größe, wenn auch bestimmte Werte (wie 1/3, 1/2, 2/3, 1 usw.) gehäuft auftreten. Lognormalverteilungen der Publikationsproduktivität findet man aber auch, wenn man die ganzzahlige Werte von

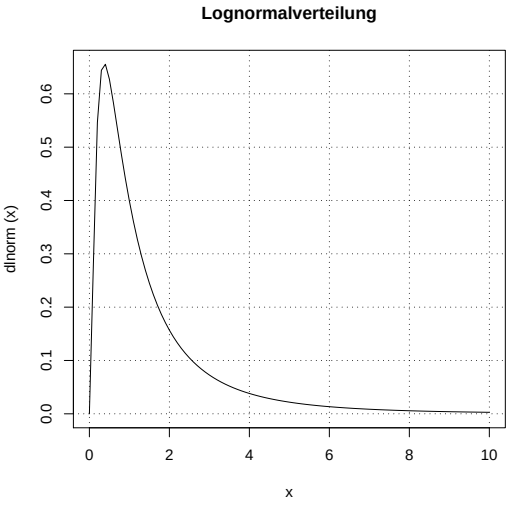


Abbildung 2.6: Dichtefunktion der Lognormalverteilung mit $\mu = 0$ und $\sigma = 1$

Artikelzahlen eines Jahrgangs von im SCI erfassten Zeitschriften betrachtet. Bradford fand bei Einschränkung auf ein Forschungsgebiet, dass die größte Zahl von Zeitschriften nur einen Aufsatz zum Thema beiträgt. Lässt man jedoch diese Einschränkung fallen, so bleibt die Verteilung der Journale nach Artikelzahlen zwar schief, aber ihr Modus[20] liegt nicht mehr bei einem Artikel (Havemann, Heinz und Wagner-Döbler 2005, Fig. 1(a), S. 5).[21]

Auch in der Ökonomie wurden lognormale Größenverteilungen – insbesondere von Firmen – nachgewiesen. Sie heißen dort nach ihrem Entdecker, dem französischem Ökonomen Robert Gibrat, der auch ein Modell zu Erklärung ihres häufigen Auftretens entwickelt hat (s. Abschnitt 4.4).

2.5 Zitationsverteilungen

Weil Forschung kollektiv und kumulativ Probleme löst und so neues Wissen schafft, das vor allem in Zeitschriften publik gemacht wird, müssen Autoren ihre Publikationen mit Referenzen auf vorhandenes Wissen versehen. Indem sie sich auf Aufsätze und Bücher beziehen, können Autoren erstens das in den zitierten Quellen publizierte Wissen verwenden und damit ihre Publikationen kurz halten. Zweitens sind sie nur durch Abgrenzung von vorhandenem Wissen in

[17] Addiert man diese für Mitarbeiter einer Gruppe auf, so gehen alle Artikel, an denen n Gruppenmitglieder mitgewirkt haben, n-fach in die Summe ein (vgl. auch Abschnitt 5.1).

[18] Gedruckte Vorläufer der INSPEC-Datenbank für Physik, Elektronik und Computing.

[19] Eine ähnliche Analyse führten 35 Jahre später Burrell und Rousseau (1995) durch.

[20] der häufigste Wert

[21] Die Abbildung ist auf der x-Achse logarithmisch skaliert, so dass sich für die Lognormalverteilung die Gauß'sche Glockenkurve ergibt.

der Lage, die Neuheit ihrer Forschungsergebnisse nachzuweisen. Einschlägige Ergebnisse anderer zu ignorieren ist unredlich, wenn damit unbegründet Priorität der eigenen Resultate suggeriert wird. Zumindest zeugt es von mangelnder Kenntnis der Literatur. Drittens wird in einführenden Bemerkungen zum Gegenstand eines Artikels gern ein Überblick über die Literatur zum Thema gegeben. Autoren können damit ihre Sicht des Forschungsgebietes propagieren. Dabei werden auch Quellen zitiert, deren Inhalte nicht direkt verwendet werden und von denen man sich auch nicht abgrenzen muss. Je breiter der Überblick gerät, um so mehr Literaturstellen werden erwähnt. *Reviews* dienen allein (oder wenigstens hauptsächlich) der zusammenfassenden Darstellung bereits publizierter Ergebnisse und sind damit nach den Originalmitteilungen die nächste Stufe im Sedimentationsprozess wissenschaftlichen Wissens, an dessen Ende das Lehrbuch steht.

Durch die Praxis des Zitierens wird die wissenschaftliche Literatur zu einem ständig wachsenden Netzwerk, in dem sich die neuen Publikationen durch ihre Referenzen auf vorhandene fachlich verorten. Diese Sicht auf Zeitschriftenaufsätze hat Derek J. de Solla Price bereits 1965 in seinem berühmten Aufsatz mit dem Titel *Networks of Scientific Papers* publik gemacht.

Entlang der zitierten Quellen von Publikationen haben schon immer Forscher für sie relevante frühere Literatur erschlossen, durch einen Index von Zitierungen kann die Recherche auch zu späteren, die Startpublikation zitierenden Aufsätzen führen. Diese Möglichkeit entstand zum ersten Mal in größerem Maßstab durch den fächerübergreifenden *Science Citation Index* (SCI) von Eugene Garfield in den sechziger Jahren des 20. Jahrhunderts. Durch den SCI wurde das durch Zitationen induzierte Netzwerk wissenschaftlicher Aufsätze bibliometrisch analysierbar. Im Kapitel zu bibliometrischen Netzwerken wird näher ausgeführt, wie fruchtbar Methoden der Netzwerkanalyse dabei sind. Hier, in diesem Abschnitt, wird vorerst von den Netzwerkeigenschaften fast vollkommen abgesehen, indem nur die Zitationszahlen von Artikeln und Büchern in die Analyse eingehen.[22]

Auf der Basis von Zitationszahlen können bibliometrische Indikatoren gebildet werden, die anzeigen, wie Aufsätze einer Bibliographie in der Forschung genutzt und beachtet werden (s. Abschnitt 5.2, S. 48). Weiterhin kann anhand von Zeitreihen die Alterung von Literatur untersucht werden (s. Abschnitt 2.6, S. 21). Hier sollen im Unterschied zu diesen beiden Zweigen der Zitationsanalyse (*citation analysis*) erst einmal nur Verteilungen von Artikeln nach der Zahl ihrer Zitationen betrachtet werden, weil sie mit den bisher vorgestellten bibliometrischen Verteilungen eins gemeinsam haben: sie sind schief.

Viele Artikel einer Fachbibliographie werden selten oder gar nicht zitiert und nur wenige erzielen größere Beachtung und Verwendung, die sich in hohen Zitationszahlen äußert. D. J. de Solla Price (1965) stellte anhand einer Garfield'schen Analyse des SCI von 1961 gerade die Schiefe der Zitationsverteilung heraus. Er berichtet, dass asymptotisch (für große Zitationszahlen) die Zahl der zitierten Artikel nach einem Potenzgesetz falle, dessen Exponent zwischen 2.5 und 3 liege.

Eine Zitationsanalyse der Genetik durch den indischen Wissenschaftler S. Naranan (1971) ergab ebenfalls eine nach einem Potenzgesetz fallende Verteilung der Artikel nach Zitationszahlen.

Einige Jahre später hat der Physiker George Magyar (1974) ein spezielles Forschungsfeld der Physik, das der Farbstofflaser (*dye lasers*), bibliometrisch analysiert. Er hat dabei auch die Verteilung von Artikeln nach ihrer Zitationszahl bestimmt, wobei er Selbstzitierungen und Zitierungen in Review-Artikeln ausschloss. Seine Untersuchung der Zitationen beschränkte er auf englischsprachige Artikel, weil andere nicht oder spät international wahrgenommen und zitiert werden.[23] Wenn man die Verteilung der Artikel nach der Zitationszahl in einem doppeltlogarithmischen Diagramm darstellt, wird die Schiefe der Verteilung deutlich: auch sie folgt (für Zitationszahlen 0 bis 14) annähernd einem Potenzgesetz (s. Abb. 2.7, S. 21).

Da $\log(0) = -\infty$, ist es angebracht, nicht den Logarithmus der Zitationszahl im Diagramm zu verwenden, sondern zur Zitationszahl 1 zu addieren. Dadurch kann auch die Zahl der nicht zitierten Artikel in die Analyse einbezogen werden. Oft wird dieses Vorgehen mit dem Argument gestützt, dass die Publikation von Forschungsresultaten einer ersten Zitierung gleichkommt.

Ermittelt man (wie Lotka, s. Abschnitt 2.1) mittels linearer Regression der logarithmierten Werte die am besten den Daten angepasste Potenzfunktion, so erhält man für den Exponenten $\alpha = 1.37 \pm 0.10$. Der Anteil der Variation der Artikelzahlen, der durch das lineare Modell

[22] In der Netzwerkanalyse entspricht die Zitationszahl dem *in-degree*, der Zahl der in einen Knoten eingehenden Kanten.

[23] Außerdem schloss er die im letzten Jahr seiner Bibliographie (1972) publizierten Artikel aus, weil sie keine genügende Chance hatten, bereits zitiert zu werden.

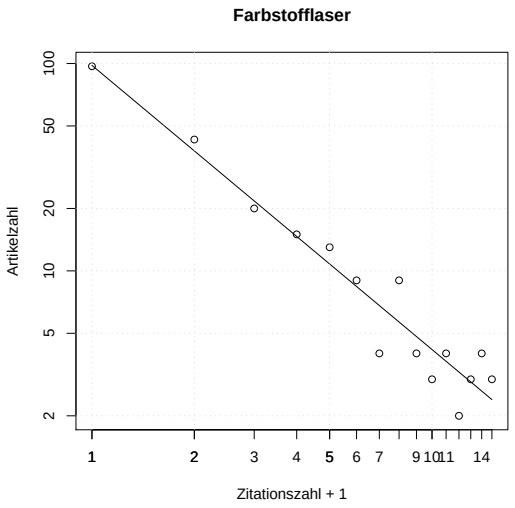


Abbildung 2.7: Diagramm der Verteilung von 233 Artikeln nach der Zahl ihrer Zitationen im Forschungsgebiet Farbstofflaser 1966–1971 in doppelt-logarithmischer Darstellung. Die Gerade entspricht einer Potenzfunktion. 18 weitere Artikel haben 16 und mehr Zitationen. Daten von Magyar (1974).

der logarithmierten Werte erklärt wird, beträgt 93% ($R^2 = .93$).

Magyar benötigte für seine Untersuchung nicht den SCI, er konnte sich bei der Ermittlung der Zitationszahlen auf die in den Aufsätzen seiner speziellen Fachbibliographie zitierten Referenzen beschränken. Die Zitierungen innerhalb eines Forschungsfeldes stellen zum Einen den größten Anteil der frühen Zitierungen (um die es hier geht), andererseits sind sie auch für die Charakterisierung des Feldes relevanter als Zitierungen von außen.

In seinem hochzitierten Aufsatz *The Skewness of Science* ging der norwegische Biologe Per O. Seglen (1992) der Frage nach, ob nicht nur der ganze SCI (wie Price berichtet) oder genügend große Fachbibliographien (wie die von Naranan und von Magyar) Zitationsverteilungen aufweisen, welche durch eine Potenzfunktion beschrieben werden können, sondern ob dies auch für einzelne Zeitschriften oder für Bibliographien von einzelnen Autoren gilt. Sollte nicht die Zitationsverteilung in diesen Fällen weniger schief sein? Er fand, dass auch hier sehr schiefe Verteilungen vorliegen, die aber nicht mit einer Potenzfunktion sondern eher exponentiell abfallen.

Die Artikel einer Bibliographie haben zu einem definierten Zeitpunkt alle eine bestimmte Zahl von zitierenden Aufsätzen, welche wiederum Elemente einer Bibliographie sind. Diese *zitierende* Bibliographie ist naturgemäß zeitlich begrenzt: das maximale *Zitationsfenster* erstreckt sich von der Publikation des zitierten Artikels bis zum Zeitpunkt der Datenerfassung für die Zitationsanalyse. Oft wird das Zitationsfenster aber eingeschränkt, um Zitationszahlen verschieden alter Artikel besser vergleichbar zu machen. Als *Publikationsfenster* wird bei der Zitationsanalyse die von der zu untersuchenden Bibliographie zitierter Artikel abgedeckte Zeitspanne bezeichnet.[24]

Für die Bibliographie der zitierenden Artikel benötigt man einen Zitationsindex, der immer nur eine Auswahl von Publikationsquellen (meist Journalen) erfasst.[25] Auch im SCI sind nicht alle naturwissenschaftlichen Zeitschriften indiziert.

Nicht nur im Zitationsnetzwerk von wissenschaftlichen Artikeln findet man Verteilungen von eingehenden Links (Zitationen), die Potenzgesetzen folgen, auch im *World Wide Web* (WWW) wurde dieses Phänomen nachgewiesen. Dort sind Weblinks auf andere Seiten das, was bei Artikeln die Zitierungen sind. Hier wie dort werden viele eingehende Links als Zeichen für Aufmerksamkeit durch die *community* gewertet (vgl. Abschnitt 5.2).

Zusammenfassung: Auch Zitationszahlen der Artikel in Fachgebieten sind schief verteilt, und zwar ebenfalls annähernd nach einem Potenzgesetz (analog zur Lotka-Verteilung wissenschaftlicher Produktivität, s. Abschnitt 2.1).

2.6 Alterung und Wachstum von Literatur

Originalmitteilungen von Forschungsergebnissen in wissenschaftlichen Publikationen enthalten neues Wissen, auf das in nachfolgenden Texten Bezug genommen wird, um es zu diskutieren, sich thematisch von ihm abzugrenzen oder um es für Forschung und Entwicklung zu verwenden (vgl. Abschnitt 2.5, S. 19).

Im Forschungsprozess entsteht ständig neues Wissen, so dass sich der Wissensschatz der Menschheit laufend erweitert. Er verbessert sich aber auch – in dem Sinne, dass Messergebnisse genauer, Methoden effektiver, Beweise einfacher werden usw. Es wird also Wissen nicht nur akkumuliert, sondern auch durch neues ersetzt oder

[24] Magyars zitierende Bibliographie ist bis auf das letzte Publikationsjahr identisch mit der zitierten: das Publikationsfenster erstreckt sich von 1966 bis 1971, das Zitationsfenster von 1966 bis 1972.

[25] Magyar erstellte seinen Zitationsindex selber aus den Referenzen seiner Fachbibliographie.

einfach nur negiert, denn in Forschungspublikationen aufgestellte Behauptungen können sich auch schlicht als falsch herausstellen. In diesem Sinne kann Wissen veralten: es wird nicht mehr gebraucht.

Auch ein Paradigmenwechsel in einem Forschungsgebiet kann Wissen veralten lassen. Der von dem Wissenschaftshistoriker Thomas S. Kuhn (1962) eingeführte Begriff des wissenschaftlichen Paradigmas wird von ihm in seinem Buch *The Structure of Scientific Revolutions* an Beispielen erläutert. Dabei wird jedoch deutlich, dass ein neues Paradigma meist das unter seinem Vorgänger erzeugte Wissen in sein Erklärungsgebäude mit aufnimmt, so wie Einsteins spezielle Relativitätstheorie Newtons klassische Mechanik als Grenzfall enthält.

Forschungsergebnisse veralten aber auch, weil der Forschungsprozess, der unvorhersehbares neues Wissen erzeugt, darum selber unvorhersehbar ist. Hoffnungsvolle Versuche ein Problem zu lösen, können in einer Sackgasse enden. Der Durchbruch gelingt anderswo. Das macht die Ergebnisse der misslungenen Versuche obsolet, zumindest zeitweilig, denn neue Methoden können auch das Forschen in der vorher aufgegebenen Richtung wieder lohnenswert machen.

Publikationen, die nur veraltetes Wissen enthalten, werden nicht mehr gebraucht. So veralten auch sie. Der Gebrauch von Publikationen kann aber auch dadurch zurückgehen, dass die in ihnen mitgeteilten Forschungsresultate in Übersichtsartikel, Monographien und Lehrbücher aufgenommen werden. Neulinge in einem Forschungsgebiet brauchen nicht alle Originalmitteilungen zu studieren, sondern können mit Hilfe didaktisch aufbereiteter, aggregierter und systematisierter Darstellungen des kumulierten Wissens die historische Entwicklung im Schnelldurchgang durchlaufen.

Der Gebrauch von Literatur zeigt sich in Bibliotheken durch Entleihung, im Internet durch das Herunterladen und in der Literatur selber durch Zitierung von Quellen. Um bibliometrisch Alterung von Literatur zu analysieren, muss die Rezeptionsgeschichte der untersuchten Bibliographie anhand von Zeitreihen von Zitationsdaten nachvollzogen werden. Dazu wird ein Zitationsindex benötigt (vgl. Abschnitt 1.2, S. 8). Will man die Alterung der Literatur eines Fachgebiets oder einer Zeitschrift charakterisieren, behilft man sich aber oft mit einer retrospektiven Analyse, die ohne Zitationsdatenbank auskommt, indem man (meist jahresweise) die zeitliche Verteilung der in einem Jahrgang zitierten Quellen (Referenzen) der Publikationen im Fachgebiet untersucht.[26] Altert die Literatur in einem Gebiet vergleichsweise schnell, werden die zitierten Referenzen hier im Mittel jünger sein als in anderen Gebieten.

Will man die Altersverteilung der in Artikeln eines Jahrgangs zitierten Quellen untersuchen, braucht man nur zu zählen, wie viele der zitierten Quellen im selben Jahr, im Jahr davor usw. publiziert worden sind. Die Altersverteilung der Quellen ist – wenn man das Publikationsdatum nur auf ein Jahr genau erfasst – annähernd gegeben durch die Verteilung der Quellen nach ihrem Publikationsjahr.[27]

Ergebnisse einer fachübergreifenden Analyse dieser Art wurden zuerst von Derek J. de Solla Price (1965) in dem bereits erwähnten Aufsatz vorgestellt (vgl. Abschnitt 2.5, S. 19). Er stützte sich dabei auf die im *Science Citation Index* (SCI) von 1961 erfassten Referenzen. Die ca. 15 Millionen Referenzen im SCI 1998 hat der Leidener Bibliometriker Anthony F. J. van Raan (2000) in gleicher Weise nach ihrem Publikationsjahr klassifiziert. Abbildung 2.8 zeigt, wie die Referenzen ab 1800 im SCI 1998 auf die Jahre verteilt sind. Wie Price und van Raan habe ich die Zahl der Referenzen logarithmisch skaliert. Dadurch wird zweierlei deutlich: Erstens liegen für einige Abschnitte auf der Zeitachse (z. B. 1950–1985) die Logarithmen der Referenzenzahlen nahe von Geraden. Die zeitliche Verteilung wird also für solche Abschnitte annähernd durch eine Exponentialfunktion beschrieben. [28]

Zweitens sind markante Einbrüche für die Jahre der Weltkriege und der Nachkriegszeiten zu beobachten, deren Ursache natürlich die Störung der Publikationstätigkeit durch den Krieg ist. Daraus wird sofort deutlich, dass die zeitliche Verteilung der Referenzen ganz stark von der Zahl der zitierbaren Quellen abhängt, welche im

[26] Retrospektive Analysen dieser Art werden – wenig suggestiv – auch als *synchrone* Alterungsstudien bezeichnet und Analysen der Rezeption eines Satzes von Publikationen als *diachrone*.

[27] Nur annähernd, weil man bei dieser Prozedur überlappende Altersintervalle von zwei Jahren erhält: z. B. können Vorjahrespublikationen am 31. Dezember erschienen sein und am 1. Januar zitiert werden, sind dann also faktisch null Jahre alt, oder es werden – im anderen Extrem – am 1. Januar des Vorjahres publizierte Artikel erst am Jahresende zitiert und sind dann nahezu zwei Jahre alt.

[28] Das ist der Grund für die Wahl der logarithmischen Skala; durch sie werden Exponentialfunktionen zu Geraden: $\log y = \log(a \exp \frac{t}{T}) = \log a + \frac{t}{T}$. Wird statt der Zahl von Referenzen ihr Alter logarithmiert, erhält man eine glockenförmiges Diagramm: Altersverteilungen von Referenzen sind annähernd lognormal, wie Matricciani (1991) und Egghe und Rao (1992) herausgefunden haben. Dem entspricht die Vorstellung, dass beim Alter nicht Differenzen sondern Verhältnisse wichtig sind. Altersdifferenzen verringern sich mit zunehmendem Alter.

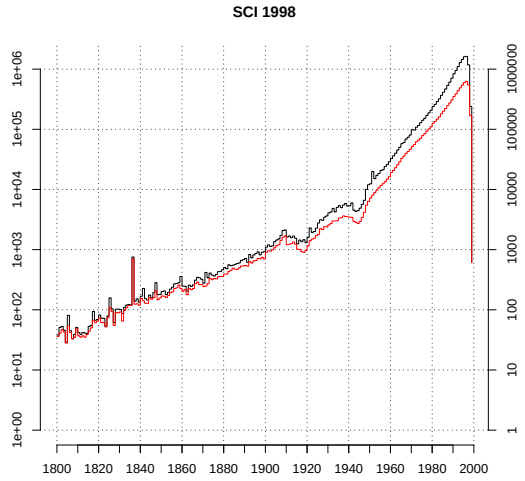


Abbildung 2.8: Verteilung der Referenzen im SCI 1998 nach ihrem Publikationsjahr (Quelle: CD-ROM-Edition SCI 1998). Die schwarze Treppenkurve entspricht der Zitationszahl der Referenzen, die rote der Zahl der unterschiedlichen zitierten Quellen.

jeweiligen Jahr publiziert wurden. Die größere Zahl jüngerer Referenzen ist vor allem dadurch erklärbar, dass die wissenschaftliche Literatur über die Jahrhunderte stark angewachsen ist und weiter stark wächst.[29]

Auch dieses Phänomen wurde zuerst von de Solla Price (1963) quantitativ untersucht. In seinem berühmten Buch *Little Science, Big Science* stellte er (u. a. anhand der Zahl wissenschaftlicher Zeitschriften seit 1667, als die ersten beiden Journale – in London und in Paris – gegründet wurden) ein exponentielles Wachstum der Wissenschaft fest:

> Unser Ausgangspunkt sind empirische statistische Aussagen über verschiedene Gebiete und Aspekte der Wissenschaften. Sie zeigen mit eindrucksvoller Konsistenz und Regelmäßigkeit, daß bei jeder vernünftigen Meßweise das normale Wachstum beliebiger genügend großer Teilstücke der Wissenschaft exponentiell erfolgt. Das heißt, Wissenschaft wächst wie ein Kapital mit Zinseszins, sie multipliziert sich in gleichen Zeitintervallen mit dem gleichen Faktor. Mathematisch folgt exponentielles Wachstum aus der einfachen Bedingung, daß zu jeder Zeit die Wachstumsrate proportional der schon erreichten Größe ist – je größer ein Ding, um so schneller sein Wachstum. In dieser Hinsicht besteht Übereinstimmung mit dem allgemeinen Naturgesetz, das das Wachstum der Weltbevölkerung, der Bevölkerung eines bestimmten Landes, der Zahl der Fruchtfliegen in einer Flaschenkolonie regiert, oder das Wachstum des Eisenbahnnetzes zu Beginn der industriellen Revolution beschreibt. [S. 16 der dt. Ausgabe]

Die Abbildung spiegelt also auch dieses starke Wachstum der Wissenschaft wieder. Price und van Raan deuten vor allem den steilen Anstieg der Referenzenzahlen, der jeweils ca. 15 Jahre vor dem Untersuchungsjahrgang beginnt, als Ausdruck der Alterung von Literatur. Für den Anstieg davor ist vorrangig das Wachstum der Literatur verantwortlich. Für diese Deutung spricht, dass die 15-Jahresperioden bei allen untersuchten Jahrgängen auftreten, nicht nur 1961 und 1998, auch bei weiteren von van Raan analysierten Jahren.

Retrospektive Analysen der Verteilung der Publikationsjahre der in einem Jahrgang zitierten Quellen reflektieren also immer beides: Alterung und Wachstum der Literatur. Eine prospektive Zitationsanalyse, d. h. die quantitative Beschreibung der Rezeptionsgeschichte von Literatur anhand der zeitlichen Verteilung der sie zitierenden Artikel, spiegelt dagegen allein deren Gebrauch. Die zeitliche Verteilung der Zitationen ist also allein Ausdruck von Literaturalterung im oben definierten Sinne.

Der Gebrauch der Literatur eines Fachgebiets ist aber natürlicherweise von dessen Wachstum abhängig. In einem Gebiet mit stagnierenden jährlichen Publikationszahlen nimmt der Gebrauch der Literatur eines Jahrgangs sicher schneller ab als in einem stark expandierenden Gebiet. Es wäre aber verfehlt, diesen Einfluss des Wachstums auf die Alterung irgendwie herausrechnen zu wollen, um so etwa die ‘wahre’ Alterung zu bestimmen. Was man jedoch herausrechnen muss, sind Erweiterungen der Zahl von Zeitschriften, die in dem für die Analyse benutzten Zitationsindex erfasst werden. Bibliographische Datenbanken erfassen im allgemeinen nicht sofort jede neu gegründete Zeitschrift; die in den Artikeln dieser Zeitschrift gemachten Zitierungen gehen damit also verloren. Wird dann die Zeitschrift indexiert, können sich von einem Jahr zum nächsten sprunghaft die ermittelten Zitierungszahlen der Artikel einer einschlägigen Bibliographie erhöhen. Durch einen festen Zeitschriftensatz verliert man wiederum einen Teil der Zitierungen.

[29] Die Erklärung der Piks in den Jahren 1836, 1951 und 1970 findet man bei van Raan (2000).

Bei Alterungsstudien kann man sowohl bei den zitierten Quellen als auch bei den zitierenden Artikeln differenzieren: einmal nach dem Dokumenttyp (Überblicksartikel oder Originalmitteilungen), dann nach der Zeitschrift (Zitierungen in der gleichen Zeitschrift oder einer anderen) und auch nach der Zitierungszahl (hochzitierte Quellen altern oft langsamer). Es gibt jedoch bislang nur wenige Analysen der Rezeptionsgeschichte von Literatur anhand der zeitlichen Verteilung der sie zitierenden Artikel.

Lange Zeit bestand die Hoffnung, Alterungsstudien könnten Bibliotheken bei Entscheidungen über die Verlagerung von Zeitschriftenbänden ins Magazin helfen. Eine solche direkte praktische Anwendung bibliometrischer Analysen in Bibliotheken fand aber kaum statt. Elektronische Speicherung von Literatur löst das Platzproblem von Bibliotheken. Die Digitale Bibliothek könnte aber irgendwann vor einem ähnlichen Problem stehen. Die Zahl direkt über das Netz abgreifbarer wissenschaftlicher Dokumente wächst ständig und wahrscheinlich exponentiell. Es könnte eines Tages unökonomisch werden, eine Unmenge von wenig oder gar nicht genutzten Dateien auf den Internet-Servern für das sofortige Herunterladen vorzuhalten.

Kapitel 3

Bibliometrische Netzwerke

3.1 Zitationsnetzwerke von Artikeln

Aufsätze in wissenschaftlichen Zeitschriften verweisen auf frühere Aufsätze und bilden dadurch ein Netz. Diese Sicht auf den Strom von Zeitschriftenliteratur hat schon vor mehr als 40 Jahren Derek J. de Solla Price (1965) propagiert (vgl. Abschnitt 2.5, S. 19). Artikel als Knoten und Zitierungen als Kanten eines Netzwerk-Graphen aufzufassen, erlaubt es, für bibliometrische Zwecke Begriffe und Methoden zu verwenden, wie sie für die Analyse sozialer Netzwerke (*Social Networks Analysis*, kurz SNA) entwickelt wurden. In größerem Ausmaß geschah dies aber erst nach dem Erscheinen des zweiten großen Informationsnetzwerkes, des *World Wide Web* (WWW), in dem *pages* die Knoten und *links* die Kanten bilden. Das Web regte Informatiker und statistische Physiker an, die für kleinere, empirisch erfassbare soziale Netze entwickelte SNA mit statistischen Methoden anzureichern, um mit dessen weitaus größeren Datenmengen fertig zu werden.

Beide Netze, das der Journalartikel wie auch das Web, erhalten ständig neue Knoten, die mit gerichteten Kanten auf schon vorhandene Knoten deuten. Während aber *Webpages* verändert werden können (oder auch wieder gänzlich gelöscht werden), bleibt ein Zeitschriftenaufsatz als Dokument nach der Publikation unverändert. Korrekturen können nur als Errata nachgetragen werden. In Webseiten können auch nachträglich Links zu späteren Seiten eingebaut werden, im Zitationsnetzwerk hingegen gibt es eine zeitliche Ordnung. Diese ist zwar nicht streng, denn Autoren können wegen der Langwierigkeit des Publikationsprozesses auch von noch nicht erschienenen Texten Kenntnis haben und sie zitieren, aber bei der Analyse von Zitationsnetzwerken wird von diesen Fällen meist abgesehen.

Das heute riesige Netzwerk wissenschaftlicher Zeitschriftenartikel begann sich herauszubilden, als das Verweisen auf ältere Publikationen übliche Praxis wurde. Ich empfinde es als hilfreich, sich das Wachstum des totalen Zitationsnetzwerks von Artikeln räumlich in Form einer wachsenden Kugel vorzustellen, welche ständig neue Jahresringe ansetzt, in denen sich die Artikel durch die von ihnen zitierten Quellen selbst verorten. Geographie und Geologie des wachsenden Globus wissenschaftlicher Literatur festzustellen, ist ein bislang unerreichtes Ziel der Zitationsanalyse. Versuche der Kartographierung jeweils eines Jahrgangs werden im Abschnitt 3.5 (S. 32) diskutiert. Erfolgreiche Entwicklungsstränge der Forschung können mit der von Eugene Garfield *et al.* entwickelten *HistCite*-Methode extrahiert werden, bei der überschaubare Netzwerke hochzitierter Aufsätze Pfade wissenschaftlicher Erkenntnis visualisieren (Garfield, Pudovkin und Istomin 2003).

Wenn nur auf Basis der durch die zitierten Quellen gegebenen Information der fachliche Ort einer Publikation bestimmt wird, verzichtet man auf alle anderen Angaben im Dokument, die für die Ortsbestimmung ebenfalls wesentlich sein können. Eine Rekonstruktion der Wissenschaftsgeschichte kann sich deshalb nicht nur auf die Analyse von Zitationsnetzwerken stützen. Zitationsanalyse hat aber den Vorteil, den mathematischen Kalkül der SNA nutzen zu können und so zu Erkenntnissen zu gelangen, zu denen hermeneutische Historiographie allein nicht fähig ist. Standardmethoden des *information retrieval* (IR) nutzen darüber hinaus die textliche Ähnlichkeit von Dokumenten, um den Nutzern zu einem gefundenen Text weitere für sie relevante anzuzeigen. Sie wurden schon oft auch für bibliometrische Zwecke eingesetzt. Das so genannte Vektormodell des IR wird weiter unten netzwerktheoretisch eingeführt (s. Abschnitt 3.6, S. 35).

Beziehungen (Kanten, Links) in einem Netzwerk gleichartiger Knoten können durch seine quadratische *Adjazenzmatrix* A mathematisch erfasst werden, deren Elemente $a_{ij} \geq 0$ von Null

verschieden sind, wenn Knoten i eine Beziehung zu Knoten j hat.[1] Wenn nicht zwischen Beziehungen verschiedener Stärke unterschieden wird, nimmt a_{ij} den Wert 1 an, falls die Beziehung vorliegt.

Derek J. de Solla Price hat in dem oben erwähnten Aufsatz die Adjazenzmatrix der Zitierungen zwischen den Artikeln einer abgeschlossenen Bibliographie über *N-rays* angegeben, indem er die Einsen durch Punkte symbolisierte und die Plätze der Nullen weiß ließ (in seiner Abb. 6, S. 514).[2] Er ordnete dafür die Artikel in der zeitlichen Reihenfolge ihres Erscheinens und ließ alle Zitierungen von Quellen außerhalb der Bibliographie weg. Die Adjazenzmatrix der ersten zwölf Artikel ist $A =$

$$= \begin{pmatrix} 0&0&0&0&0&0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0&0&0&0&0 \\ 1&1&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&1&1&0&0&0 \\ 0&0&0&0&0&0&0&1&1&0&0&0 \\ 0&0&0&0&0&0&0&1&1&1&0&0 \end{pmatrix}.$$

Weil zukünftige Artikel nicht zitiert werden können, kommen in den Zeilen dieser Matrix Einsen nur vor der Hauptdiagonale vor. Wegen der zeitlichen Ordnung des Zitationsnetzwerkes hat seine Adjazenzmatrix die Gestalt eines Dreiecks: oberhalb und auf der Hauptdiagonalen finden sich nur Nullen ($a_{ij} = 0, \forall j \geq i$).[3]

Der Graph des Netzwerks der ersten zwölf Artikel in Abbildung 3.1 zerfällt in zwei Teilgraphen. Jeder steht für unabhängig gewonnene Resultate, die erst später im ersten Überblicksartikel zu *N-rays* (Nr. 75 der Bibliographie) als zusammengehörig interpretiert wurden.[4]

Die Adjazenzmatrix A des Zitationsnetzwerkes kann zur Modellierung des Verhaltens eines Lesers benutzt werden, der sich anhand der zitierten Quellen von Artikel zu Artikel bewegt. Der Leser lese zur Anfangszeit $t = 0$ z. B. Artikel Nr. 12. Dann wird er durch einen Spaltenvektor $\vec{r}(0)$ beschrieben, der elf Nullen enthält und eine Eins als zwölfte Komponente. Multipliziert man ihn von links mit der Transponierten von

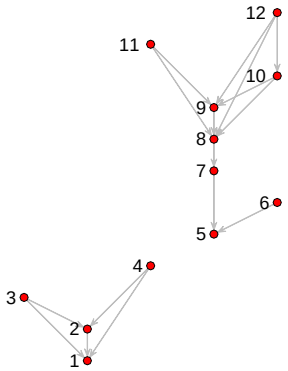


Abbildung 3.1: Zeitlich geordneter Graph des Zitationsnetzwerks zwischen den ersten zwölf Artikeln über *N-rays*. Datenquelle: de Solla Price (1965)

A, findet man den Leser bei den Artikeln Nr. 8, 9 und 10. Im nächsten Schritt wandert er durch die Vorschrift $\vec{r}(t+1) = A^{\mathrm{T}}\vec{r}(t)$ (oder kürzer $\vec{r} \leftarrow A^{\mathrm{T}}\vec{r}$) von dort zu Nr. 7, 8 und 9, usw.:

$$\vec{r}(0) = \begin{pmatrix} 0\\0\\0\\0\\0\\0\\0\\0\\0\\0\\0\\1 \end{pmatrix}, \vec{r}(1) = \begin{pmatrix} 0\\0\\0\\0\\0\\0\\0\\1\\1\\1\\0\\0 \end{pmatrix}, \vec{r}(2) = \begin{pmatrix} 0\\0\\0\\0\\0\\0\\1\\2\\1\\0\\0\\0 \end{pmatrix}, \ldots$$

Dieses Modell eines Lesers mutet hier noch wie eine Spielerei an, was es ja auch sein soll, denn spielend lernt man bekanntlich am leichtesten. Im folgenden Abschnitt zeige ich, wie dieses Modell direkt zu dem des *Random Surfers* führt, das Brin und Page (1998) ihrem *PageRank*-Algorithmus zu Grunde legten, welcher den Erfolg von *Google* nicht unwesentlich mit bewirkte. Als Vorarbeit dazu wollen wir unser Modell zu dem des *Random Readers* verfeinern, indem wir die Komponenten der Vektoren $\vec{r}$ jeweils auf die Summe 1 normieren: $\vec{R} = \vec{r}/\sum r_i \equiv \vec{r}/r_+$. Dadurch werden beispielsweise zum Zeitpunkt $t = 2$ die von Null verschiedenen Elemente des Leser-Vektors zu $R_7(2) = 1/4$, $R_8(2) = 1/2$ und $R_9(2) = 1/4$. Durch die Normierung auf 1 sind diese Anteile als Wahrscheinlichkeit interpretierbar, und zwar als die Wahrscheinlichkeit mit der

[1] Adjazent [lateinisch] Anrainer, Grenznachbar

[2] *N-rays* stellten sich als fiktiv heraus, deswegen kann die Bibliographie als abgeschlossen gelten.

[3] Zeitlich geordnete Netzwerke nennt man auch *azyklisch*, weil man in ihnen nicht entlang der gerichteten Links in einem Kreislauf zum Ausgangspunkt zurückkehren kann.

[4] was sich als Kozitation äußert (vgl. Abschnitt 3.5, S. 32)

wir den Leser zur Zeit $t = 2$ bei Artikel Nr. 7, 8, oder 9 antreffen, wenn er nach der Lektüre eines Textes jeweils zufällig eine zitierte Quelle wählt. Bei Artikel 8 treffen wir ihn mit doppelter Chance, weil er zu ihm sowohl über 9 als auch über 10 gelangen kann. Das macht deutlich, warum der Leser durch die Normierung nun als *Random Reader* bezeichnet werden kann.

Mit Hilfe von Zitationsindizes können Leser heute nicht nur retrospektiv Artikel über die Referenzenlisten zitierter Quellen finden, sondern auch zeitlich vorwärts im Zitationsnetzwerk von Artikeln navigieren.[5] Man kann diesen Prozess modellieren, indem man die Transponierte der transponierten Adjazenzmatrix verwendet, d. h. A selber,[6] weil die Spiegelung an der Hauptdiagonalen alle Pfeile des Graphen eines gerichteten Netzwerks umkehrt, wie man sich leicht klarmacht.

Die Adjazenzmatrix A gibt die direkten Wege zwischen Knoten im Netzwerk entlang der gerichteten Kanten an. Für das Modell des Lesers haben wir Potenzen von A (bzw. von A^{T}) benutzt: $A^1, A^2, A^3 \ldots$ Wenn wir z. B. $A^2 =$

$$= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 \end{pmatrix}$$

mit dem Graphen in Abbildung 3.1 (S. 26) vergleichen, sehen wir, dass die Elemente von A^2 angeben, wie viele (indirekte) Wege der Länge 2 zwischen den Knoten existieren, z. B. zwei von Knoten 12 zu Knoten 8 (zwischen anderen Knotenpaaren ist höchstens ein Weg der Länge 2 vorhanden). Allgemein gilt, dass die Matrix A^k die Zahlen der Wege der Länge k zwischen den Knoten enthält.

3.2 Zitationsnetzwerke von Journalen

Forschung konnte über die Jahrhunderte nur deshalb so stark an Umfang zunehmen, weil ständig neue Spezialgebiete entstanden, zwischen deren Fachgemeinschaften die Forschungsarbeit aufgeteilt wurde. Jede neue Fachgemeinschaft schuf sich ihr Publikationsorgan, eine Fachzeitschrift. Daneben bestanden und bestehen fachübergreifende Journale weiter fort, in welchen Forschungsresultate von allgemeineren Interesse publik gemacht werden. Aber auch eng spezialisierte Zeitschriften enthalten nicht nur Beiträge von Spezialisten des Gebietes, sondern auch Artikel aus anderen Forschungsrichtungen, welche für die Leser der jeweiligen Zeitschrift interessant sein könnten. Das gerade bewirkt, dass Literatur eines Gebiets nicht nur in dessen Kernjournalen zu finden ist, sondern breit gestreut wird, wie durch Bradfords Gesetz beschrieben (s. Abschnitt 2.3, S. 16).

Trotz dieser Beimengung fachfremder Beiträge sollten in Artikeln einer Zeitschrift andere Zeitschriften aus – im Stammbaum der Wissenschaft – benachbarten Forschungsrichtungen häufiger zitiert werden, als die von weiter entfernten. Diese plausible Annahme lag schon der Untersuchung von Gross und Gross (1927) zugrunde, die Zitierungszahlen von anderen Zeitschriften in dem allgemein-chemischen *Journal of the American Chemical Society* bestimmten, um damit Bibliothekaren Hinweise für ihre Zeitschriftenauswahl zu geben. So erhobene Zitierungszahlen von Journalen sind jedoch nicht nur von fachlicher Nähe oder Distanz beeinflusst, sondern auch einfach durch die Zahl der Artikel im zitierten Journal und durch deren Qualität. Journale ähnlicher Ausrichtung stehen in Konkurrenz um die für die weitere Forschung bedeutsamsten Artikel. Deswegen unterziehen sie die eingesandten Aufsätze einem Begutachtungsverfahren, dem *peer review*. Je größer die Reputation einer Zeitschrift, umso mehr Aufsätze bekommt sie zugesandt und umso strenger kann die Begutachtung ausfallen. So differenzieren sich wissenschaftliche Periodika nicht nur fachlich, sondern auch nach ihrem Ansehen.

Die Zitationsströme im Netzwerk von Fachzeitschriften sind also durch diese drei Faktoren beeinflusst: durch fachliche Nachbarschaft, durch die Größe der Zeitschriften und durch ihr Ansehen. Als weitere Einflussgröße kommt noch die im Gebiet übliche Zahl von zitierten Quellen pro Aufsatz hinzu.

Wenn zitierte Quellen unabhängig vom Erscheinungsjahr in die Analyse einbezogen werden, spielt das Alter der Journale eine Rolle für die Zahl von Artikeln, die zitiert werden können, und damit auch für die Zahl der Zitationen, die sie bekommen können. Bei der Konstruktion eines Zitationsnetzwerkes von Zeitschriften sollte man deshalb ein Zeitfenster für die Publikations-

[5] Durch abwechselndes Rückwärts- und Vorwärtsgehen können sich Nutzer auf Zickzacklinieren auch seitwärts im Graphen bewegen, z. B. von Artikel Nr. 3 über 2 zu 4 (s. Abschnitt 3.4, S. 31).

[6] $(A^{\mathrm{T}})^{\mathrm{T}} = A$

jahre der zitierten Quellen festlegen, in welchem alle untersuchten Zeitschriften durchgängig existiert haben.

Der Einfluss der Größe der Journale kann herausgerechnet werden, indem man die Zitationszahl jeweils auf die Zahl zitierbarer Artikel bezieht, wie Eugene Garfield es bei seiner Definition des *Journal Impact Factors* (JIF) macht (s. Abschnitt 5.2, S. 48). Um dabei auch die unterschiedliche Zitationsgewohnheiten in verschiedenen Gebieten zu berücksichtigen, haben Pinski und Narin (1976) vorgeschlagen, die Zitationszahl einer Zeitschrift nicht auf deren Artikelzahl, sondern auf die Gesamtzahl der in ihren Artikeln vergebenen Referenzen zu beziehen. Das läuft auf eine Art Import-Export-Relation hinaus, bei der auch ausgeglichen wird, dass Review-Artikel mit langen Referenzenlisten im Durchschnitt häufiger zitiert werden als Originalmitteilungen von Forschungsergebnissen. In dieser Weise konstruierte Zitationsnetzwerke von Zeitschriften spiegeln dann deren fachlichen Verwandtschaftsverhältnisse und ihre jeweilige Reputation.

Im Vergleich zu Zitationsnetzwerken von Artikeln enthalten Zeitschriftennetze natürlicherweise weitaus weniger Knoten und sind daher nicht nur übersichtlicher, sondern auch einer numerischen Analyse leichter zugänglich. Die für die Konstruktion von Zeitschriftennetzwerken nötigen Zitationszahlen werden in den *Journal Citation Reports* des *Science Citation Index* (und des *Social Sciences Citation Index*) im Web aggregiert bereitgestellt.

Als Beispiel betrachten wir verschiedene Varianten eines Netzwerks der folgenden fünf informationswissenschaftlichen Zeitschriften:

1. *Information Processing & Management,*
2. *Journal of the American Society for Information Science and Technology,*
3. *Journal of Documentation,*
4. *Journal of Information Science,*
5. *Scientometrics.*

Wir konstruieren zunächst das mit den gegenseitigen Zitationszahlen gewichtete Netz (inklusive der Selbstzitationen der Journale). Mit den Daten der *Social Sciences Edition* der *Journal Citation Reports* erhalten wir für das Zitationsfenster 2006 und das Publikationsfenster 2002-2006 folgende Adjazenzmatrix:

$$A = \begin{pmatrix} 79 & 65 & 15 & 6 & 24 \\ 42 & 182 & 11 & 15 & 44 \\ 6 & 22 & 37 & 8 & 6 \\ 20 & 26 & 13 & 30 & 11 \\ 7 & 48 & 7 & 10 & 254 \end{pmatrix}.$$

Die Matrix A enthält nur Elemente $a_{ij} \neq 0$, weil alle fünf Journale sich gegenseitig und auch sich selbst zitieren. Die Journal-Selbstzitierungen füllen die Hauptdiagonale (d. s. die Elemente a_{ii}). Im zu A gehörigen Graphen sind alle Kanten (*links*) zwischen den fünf Knoten in beiden Richtungen realisiert und ebenso die durch Schleifen darstellbaren Selbstzitierungen der Journale (*self links*). Es fällt auf, dass die Selbstzitierungen der Journale recht häufig sind. Wir quantifizieren dieses Gefühl durch den Vergleich des Netzwerks mit einem Modellnetz, in dem die Kanten zufällig verteilt sind, die Gesamtzahl von erhaltenen und vergebenen Zitierungen für jede Zeitschrift aber möglichst erhalten bleiben soll. Das erreicht man, indem eine Kante zwischen den Knoten i und j mit der Wahrscheinlichkeit $p_i q_j$ realisiert wird, wobei die Randwahrscheinlichkeiten p_i und q_j aus den entsprechenden relativen Häufigkeiten berechnet werden:

$$p_i = \frac{\sum_j a_{ij}}{\sum_{i,j} a_{ij}} \equiv \frac{a_{i+}}{a_{++}}, q_j = \frac{a_{+j}}{a_{++}}.$$

Dieses Vorgehen ist einem χ^2-Test gleichbedeutend. Für das Zufallsnetzwerk erwartet man eine Adjazenzmatrix mit den Elementen $p_i q_j a_{++}$ $(i, j = 1, \ldots, 5)$, d. h. gerundet (auf ganzzahlige Werte):

$$\begin{pmatrix} \mathit{29} & 66 & 16 & 13 & 65 \\ 46 & \mathit{102} & 25 & 21 & 101 \\ 12 & 27 & \mathit{7} & \mathit{6} & 27 \\ \mathit{16} & 35 & \mathit{8} & \mathit{7} & 34 \\ 51 & 113 & 27 & 23 & \mathit{112} \end{pmatrix}.$$

Die kursiv gedruckten Elemente der Matrix der Erwartungswerte sind kleiner als die entsprechenden Elemente von A. Es ist ersichtlich, dass vor allem die Hauptdiagonalelemente von A von den Erwartungswerten deutlich nach oben abweichen. Jede Zeitschrift ist sich selbst die nächste, was nicht nur in fachlicher Hinsicht zu interpretieren ist.[7] Als Maß der Abweichung vom Erwarteten wird beim χ^2-Test die Wahrscheinlichkeit berechnet, mit der das durch A beschriebene Netzwerk als Ergebnis des Zufallsprozesses erscheinen kann. Sie ist mit $2.2 \cdot 10^{-16}$ verschwindend gering.

Wir gehen nun – wie Pinski und Narin (1976) – zu einem Netzwerk über, wo die Zitationszahl a_{ij} von Zeitschrift j durch Zeitschrift i durch die Summe a_{j+} der Referenzenzahlen von j im Netzwerk der fünf Zeitschriften dividiert wird.

[7] Es wurden Fälle bekannt, wo Herausgeber Druck auf Autoren ausübten, um sie zum Zitieren ihrer Zeitschrift zu bewegen.

Die Adjazenzmatrix mit den Elementen $\gamma_{ij} = a_{ij}/a_{j+}$ wird dann (gerundet) zu

$$\gamma \approx \begin{pmatrix} .42 & .22 & .19 & .06 & .07 \\ .22 & .62 & .14 & .15 & .13 \\ .03 & .07 & .47 & .08 & .02 \\ .11 & .09 & .16 & .30 & .03 \\ .04 & .16 & .09 & .10 & .78 \end{pmatrix}. \quad (3.1)$$

Pinski und Narin gehen noch weiter. Sie argumentieren, dass eine Zitierung in einem Journal mit hohem Prestige mehr zählen sollte als eine in einer weniger wichtigen Zeitschrift. Da sie Prestige aber gerade in erhaltenen Zitierungen (pro zitierter Quelle) messen, gelangen sie zu einem rekursiven Begriff von Prestige, wie er für die Analyse sozialer Netzwerke seit den 1940-er Jahren diskutiert wird (Wasserman und Faust 1994). In sozialen Netzwerken kommt es nicht nur darauf an, viele Leute zu kennen, es müssen auch die richtigen sein, nämlich solche, die ihrerseits viele richtige kennen.

Wie nun diesen rekursiven Prestige-Begriff mathematisch fassen? Dazu betrachten wir einen Modellprozess der Umverteilung von Prestige für die fünf informationswissenschaftlichen Zeitschriften. Zu Beginn ($t = 0$) sollen alle fünf Journale gleiches Gewicht haben, das wir auf $w_j(0) = 1, \forall j$, festsetzen. Die Gewichte $w_i(0)$ schreiben wir in einen Spaltenvektor $\vec{w}(0)$. Analog zum Vorgehen beim Leser-Modell für das Zitationsnetzwerk von Artikeln (vgl. S. 26) multiplizieren wir $\vec{w}(0)$ von links mit der transponierten Adjazenzmatrix, die hier durch γ^{T} (s. Gleichung 3.1) gegeben ist: $\vec{w}(1) = \gamma^{\mathrm{T}}\vec{w}(0)$. Dadurch werden die Gewichte im Netzwerk umverteilt. Der Vektor $\vec{w}(1)$ enthält gerade die Reihensummen von γ^{T}, nämlich $w_j(1) = \gamma_{+j} = a_{+j}/a_{j+}$, d. h. die Import-Export-Relationen der Journale. Wir wiederholen nun die Umverteilungsprozedur immer wieder gemäß $\vec{w}(t+1) = \gamma^{\mathrm{T}}\vec{w}(t)$ (oder kürzer: $\vec{w} \leftarrow \gamma^{\mathrm{T}}\vec{w}$) und bemerken, dass sich die Gewichte $w_j(t)$ iterativ festen Grenzwerten nähern:[8]

$$\vec{w}(1) \approx \begin{pmatrix} 0.81 \\ 1.17 \\ 1.05 \\ 0.69 \\ 1.04 \end{pmatrix}, \vec{w}(\infty) \approx \begin{pmatrix} 0.69 \\ 1.21 \\ 0.93 \\ 0.59 \\ 1.14 \end{pmatrix}.$$

[8] Die Differenzen $\vec{w}(t+1) - \vec{w}(t)$ werden für $t \to \infty$ immer kleiner.

Demnach gilt für $t \to \infty$ die Gleichung $\vec{w} = \gamma^{\mathrm{T}}\vec{w}$,[9] oder komponentenweise geschrieben:

$$w_j = \sum_{i=1}^{n} \gamma_{ij} w_i \quad (3.2)$$

mit $n = 5$. Das bedeutet, dass die durch die Iteration ermittelten Gewichte eine Gleichung erfüllen, die als Ausdruck der rekursiven Prestige-Definition angesehen werden kann: Das Gewicht von Journal j ergibt sich aus den Zitationsbeziehungen zu allen anderen Journalen (und auch zu sich selbst) nach Maßgabe der Enge dieser Beziehungen multipliziert mit den Gewichten der anderen Journale.[10]

Es muss noch angemerkt werden, dass die Iterationsprozedur $\vec{w} \leftarrow \gamma^{\mathrm{T}}\vec{w}$ von mir etwas zu Unrecht als Umverteilung bezeichnet wurde. Die Summe der fünf Gewichte nach dem ersten Iterationsschritt ist $w_+(1) \approx 4.76 < n = 5$. Es geht hier also Gewicht verloren (weil die Import-Export-Relationen der großen Journale günstiger sind als die der kleinen). Dies kann durch Normierung behoben werden; für $t \to \infty$ erhalten wir: $n\vec{w}^{\mathrm{T}}/w_+ \approx (0.76, 1.33, 1.03, 0.64, 1.25)$. Mit dieser Normierung auf n wird deutlich, wer zu den Gewinnern und wer zu den Verlierern der Umverteilung gehört.

Ein etwas abgewandelter Umverteilungsprozess des Gewichtes von Journalen wurde von Nancy Geller (1978) in Anschluss an Pinski und Narin (1976) betrachtet. Ihr Ausgangspunkt ist die Theorie der Markow-Ketten, einer speziellen Klasse von stochastischen oder Zufallsprozessen, bei denen – wie bei uns – der Zustand zur Zeit $t+1$ vollkommen durch den Zustand zur Zeit t bestimmt ist.[11] Sie verwendet statt γ eine etwas anders normierte Matrix γ^* mit den Elementen $\gamma^*_{ij} = a_{ij}/a_{i+}$, deren Transponierte $\gamma^{*\mathrm{T}}$ zu den stochastischen Matrizen gehört. Diese haben die Eigenschaft, die Normierung der Vektoren auf die Summe ihrer Elemente bei Multiplikation unverändert zu lassen. Stochastische Matrizen beschreiben also echte Umverteilungen und sind deswegen für die Beschreibung von Zufallsprozessen geeignet, bei denen Wahrscheinlichkeit auf mögliche Systemzustände umverteilt

[9] Dass sich das nicht nur in unserem speziellen Fall, sondern immer so verhält, wird durch die mathematische Theorie von Eigenwertproblemen $B\vec{x}_k = \lambda_k\vec{x}_k$ garantiert. Matrizen der Bauart von γ^{T} haben einen maximalen Eigenwert $\lambda_{\max} = 1$ und die Iteration ermittelt den dazugehörigen Eigenvektor $\vec{x}_{\max}$.

[10] Pinski und Narin nennen das Prestige-Gewicht für Journale *influence weight*. Bestimmungsgleichungen dieser Art werde auch als *bootstrap relations* bezeichnet (nach dem englischen Wort für Stiefelschlaufe).

[11] Es handelt sich also um Prozesse mit kurzem Gedächtnis.

wird. Der Algorithmus von Nancy Geller ist für uns hier deshalb interessant, weil von ihm aus nur wenige Schritte nötig sind, um zum *PageRank*-Algorithmus von Google zu gelangen, wie jetzt – als Exkurs in das Gebiet des *information retrieval* – dargestellt wird.

3.3 Exkurs: *PageRank*-Algorithmus

Eine wesentliche Ursache für den Erfolg von Google ist sicher der von den Gründern Brin und Page (1998) publizierte Algorithmus für das Ranking von Webpages nach dem Grad ihrer Verlinkung mit anderen Webpages. Sie wenden dabei das gleiche SNA-Prinzip an wie Pinski und Narin (1976). Übersetzt in die Sprache des Web lautet es bei Brin und Page:

> Academic citation literature has been applied to the web, largely by counting citations or backlinks to a given page. This gives some approximation of a page's importance or quality. PageRank extends this idea by not counting links from all pages equally, and by normalizing by the number of links on a page.

Die Normierung der Adjazenzmatrix A erfolgt wie bei Geller (1978) gemäß $\gamma^*_{ij} = a_{ij}/a_{i+}$. Das Web wird hier als ungewichtetes Netzwerk modelliert, wo mehrfache Links von einer Seite zu einer anderen nur wie ein Link behandelt werden (analog zum Zitationsnetzwerk wissenschaftlicher Aufsätze). Die Matrix A ist hier also (wie dort) eine binäre (sie besteht nur aus Nullen und Einsen). Nun gibt es Webpages, die überhaupt keine Links auf andere Seiten setzen. Beim Umverteilungsprozess des *PageRank*-Gewichts entlang der Weblinks bekommen sie in jeder Runde etwas zugeteilt, was sie aber weder weitergeben noch behalten. Es wird vernichtet, weil ihre Zeilen in der Adjazenzmatrix nur aus Nullen bestehen, wodurch die Transponierte von γ^* nicht mehr stochastisch ist.

Diesen Umstand kann man beheben, indem man alle Elemente der Hauptdiagonale von A gleich eins setzt, was für alle Webpages einen *self link* annimmt. Dadurch wird die Vernichtung von *PageRank*-Gewicht gestoppt, aber die Seiten ohne *out links* häufen bei jeder Iteration mehr und mehr Gewicht an und alle Seiten *mit* Links zu anderen verlieren immer mehr. Dieses Problem besteht im Netzwerk von Journalen nicht: eine wissenschaftliche Zeitschrift, die keine Zitierungen von Artikeln in anderen Journalen enthält, gibt es nicht.

Die wesentliche Idee von Brin und Page zur Lösung dieses Problems bestand nun darin, von allen Webpages in jeder Runde einen gewissen Prozentsatz von Gewicht wegzunehmen – als eine Art Steuer – und diesen Betrag an alle gleichmäßig zu verteilen. Auch dieser Prozess konvergiert zu einer stabilen Endverteilung von Gewicht, wie man zeigen kann.

Brin und Page notierten ihren Algorithmus komponentenweise. Das *PageRank*-Gewicht $w_j(t+1)$ der Webpage j ergibt sich aus den Gewichten $w_i(t)$ aller Webpages i, die einen Link zu j gesetzt haben (bei Normierung $w_+ = N$) zu

$$\begin{aligned} w_j(t+1) &= 1-d+d\textstyle\sum_{i\to j} w_i(t)/a_{i+} \\ &= 1-d+d\textstyle\sum_{i=1}^{N} w_i(t)\gamma^*_{ij}, \end{aligned} \tag{3.3}$$

wobei $0 < d < 1$ als Dämpfungsfaktor (*damping factor*) bezeichnet wird. Wenn, als Beispiel, $d = 85\%$ ist, werden die restlichen $1-d = 15\%$ gleichmäßig auf alle Gewichte w_j verteilt und nicht über die selbstverstärkende Rückkopplung, welche auf 85% gedämpft wird.

Um zum Modell des *Random Surfers* zu kommen, gehen wir mit $\vec{p} = \vec{w}/N$ zu einem auf Eins normierten Gewichtsvektor $\vec{p}$ über ($p_+ = 1$). Seine Komponenten $p_i(t)$ geben die Wahrscheinlichkeit an, den Surfer zum Zeitpunk t auf Seite i anzutreffen, wenn dieser zu Beginn ($t = 0$) zufällig von irgendeiner Seite gestartet ist und dann bei jedem Zeittakt mit Wahrscheinlichkeit d irgendeinem Link zu einer anderen Seite gefolgt ist oder (mit Wahrscheinlichkeit $1-d$) wiederum eine von allen N Seiten zufällig als neue Startseite ausgewählt hat. Letzteres verhindert, dass der Surfer sich in Gebieten des Web verfängt, aus denen keine Links nach außen führen (auch als *spider traps* bezeichnet).

Wir können diesen Zufallsprozess kompakter beschreiben, wenn wir die (stochastische) Gleichverteilungsmatrix U (*uniform transition matrix*) einführen, deren Elemente alle gleich sind: $u_{ki} = 1/N, \forall k, i = 1, \ldots, N$. Sie entspricht einem total verlinkten Netzwerk. Es gilt

$$\sum_{i=1}^{N} u_{ki}p_i = \frac{1}{N}\sum_{i=1}^{N} p_i = \frac{1}{N}.$$[12]

Der *PageRank*-Algorithmus schreibt sich dann als

$$\vec{p} \leftarrow [(1-d)U + d\gamma^{*\mathrm{T}}]\vec{p}. \tag{3.4}$$

Dieser iterative Algorithmus konvergiert: es stellt sich ein Fließgleichgewicht von Wahrscheinlichkeit ein. Am Ende finden wir den *Random Surfer* mit Wahrscheinlichkeit $p_i(\infty)$ auf

[12] Die Gleichverteilung ist also ein Prozess ohne Gedächtnis.

Seite i, und zwar mit höherer Wahrscheinlichkeit auf gut verlinkten Seiten, die demgemäß beim Ranking der Suchergebnisse weiter oben platziert werden sollten.

3.4 Bibliographische Kopplung

Zwei Artikel nennt man mit Kessler (1963) bibliographisch gekoppelt, wenn mindestens eine zitierte Quelle in den Bibliographien beider Artikel auftaucht. Der Zitationsgraph in Abbildung 3.2 illustriert die Kopplung von Artikel 1 und 2 über die zitierte Quelle 3.

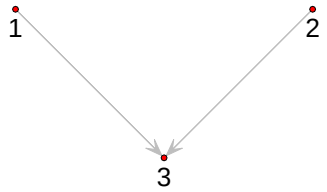


Abbildung 3.2: Bibliographische Kopplung: In den Artikeln 1 und 2 wird Artikel 3 zitiert.

Wenn man im Graphen der zwölf ersten Aufsätze zu *N-rays* (Abb. 3.1, S. 26) nach bibliographischen Kopplungen sucht, entdeckt man mehrere: Die Knoten 2 und 3 sind über Knoten 1 bibliographisch gekoppelt, wie auch Knoten 2 und 4. Die Knoten 3 und 4 sind sowohl über Knoten 1 als auch über Knoten 2 gekoppelt, Knoten 6 und 7 über 5; die Knoten 9 bis 12 sind paarweise über 8 gekoppelt, die Knoten 10 bis 12 darüber hinaus auch noch über Knoten 9.

Wenn ein Leser von einem Artikel über das Zitationsnetzwerk zu einem bibliographisch gekoppelten Artikel gelangen will, muss er zuerst im Graphen einen Schritt entlang der Pfeile gehen und dann einen gegen die Pfeilrichtung. Wir wissen, dass dieser Prozess durch die Matrix $B = AA^{\mathrm{T}}$ beschrieben wird. Ihr Element b_{ij} entsteht nach den Regeln der Matrixmultiplikation als Skalarprodukt der Zeilenvektoren von A: $b_{ij} = \sum_k a_{ik}a_{jk}$. Da A eine binäre Matrix ist, ergibt die Summierung die Zahl der übereinstimmenden Komponenten der beiden Zeilenvektoren, d. h. die Zahl gemeinsamer Quellen. Für das Zitationsnetzwerk des betrachteten Beispiels berechnen wir Matrix B zu

$$\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 & 2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 2 & 3
\end{pmatrix}.$$

Element b_{ij} gibt also an, wie viele bibliographische Kopplungen zwischen den Artikeln i und j bestehen. Anders ausgedrückt, gibt es die Zahl von Wegen der Länge 2 an, bei denen man sich von i entlang der Pfeile und zu j in Gegenrichtung bewegt (vgl. S. 27). Der Symmetrie der Kopplungsbeziehung entspricht die der Matrix: $b_{ij} = b_{ji}$, denn wegen $(AA^{\mathrm{T}})^{\mathrm{T}} = (A^{\mathrm{T}})^{\mathrm{T}}A^{\mathrm{T}} = AA^{\mathrm{T}}$ gilt $B^{\mathrm{T}} = B$.

Die Hauptdiagonale enthält die Zahlen der bibliographischen Selbstkopplungen der Artikel, nämlich die Zahlen aller ihrer Referenzen auf andere Artikel im Netzwerk.

Zitationsdatenbanken ermöglichen es Nutzern, sich im Zitationsnetzwerk zeitlich vorwärts und rückwärts, aber auch seitwärts zu bewegen. Thematisch ähnliche Artikel, die im selben Jahrgang erscheinen, sind sich zeitlich oft zu nah, als dass im späteren Artikel der frühere zitiert werden könnte. Sie verraten ihre Ähnlichkeit aber auch durch ähnliche Referenzenlisten, d. h. durch eine starke bibliographische Kopplung. Schon in der seit Ende der 1980-er Jahre vertriebenen CD-ROM-Edition des *Science Citation Index* (SCI) wurde man von einem aufgefundenen Artikel mit der Option *Related Records* auf die zwanzig mit ihm am stärksten bibliographisch gekoppelten Artikel verwiesen. Die Stärke der Kopplung zweier Artikel i und j wird hier einfach als Zahl der beiden gemeinsamen Referenzen definiert, wie sie durch das Element b_{ij} der Matrix $B = AA^{\mathbf{T}}$ gegeben ist.

Beim Beispiel der Bibliographie zu *N-rays* können wir nur Zitationsbeziehungen zwischen ihren Artikeln in die Analyse einbeziehen, obwohl in ihren Referenzenlisten sicher auch Quellen zitiert werden, die nicht zur Bibliographie gehören. Analysieren wir, zum Beispiel, einen Jahrgang des SCI, dann wählen wir keinen thematischen Ausschnitt aus dem Zitationsgraphen, sondern betrachten alle im Jahrgang des SCI erfassten Journalartikel mit allen ihren zitierten Quellen, welche auch Bücher, Patente,

Zeitungsartikel usw. sein können. Dem angemessen ist nicht eine quadratische Adjazenzmatrix A, deren Zeilen und Spalten die selben Knoten repräsentieren,[13] sondern eine Rechteckmatrix, die für jeden Artikel eine Zeile und für jede zitierte Quelle eine Spalte enthält. Nur wenige der Journalartikel des Jahrgangs werden auch bei den zitierten Quellen auftauchen. Wir gelangen so zu einem Zitationsnetzwerk von zweierlei Knoten: Artikel und Quellen. In ihm sind nur Kanten zwischen verschiedenartigen Knoten erlaubt. Netzwerke dieser Art werden auch als *bipartit* bezeichnet. Ihre Adjazenzmatrix wird Affiliationsmatrix genannt.[14]

Auch für eine rechteckige Affiliationsmatrix A mit m Zeilen (Artikeln) und n Spalten (Quellen) kann die Matrix der bibliographischen Kopplung $B = AA^{\mathbf{T}}$ berechnet werden. Matrix B ist auch in diesem Fall quadratisch und enthält für jeden der m Artikel eine Reihe und eine Spalte.

Zwei Artikel mit langen Referenzenlisten können viele gemeinsame Quellen haben; Artikel mit wenigen Referenzen sind daher tendenziell schwächer bibliographisch gekoppelt, wenn man die Stärke der Kopplung einfach nur mit der Zahl gemeinsamer Referenzen b_{ij} misst. Das legt nahe, zu einem relativen Maß der bibliographischen Kopplung überzugehen. Dieses kann am einfachsten mengentheoretisch definiert werden. In Referenzenlisten taucht jede zitierte Quelle nur einmal auf; deswegen kann man sie auch als Mengen zitierter Quellen ansehen. In der Sprache der Mengenlehre formuliert, gilt also $b_{ij} = |\mathbf{R}_i \cap \mathbf{R}_j|$, d. h. das Element b_{ij} der Matrix B ist gleich der Größe des Durchschnitts der Referenzenlisten der Artikel i und j. Als relatives Maß der Überlappung zweier Mengen steht der *Jaccard-Index* zur Verfügung:[15]

$$J_{ij} = \frac{|\mathbf{R}_i \cap \mathbf{R}_j|}{|\mathbf{R}_i \cup \mathbf{R}_j|}. \tag{3.5}$$

Der Jaccard-Index der bibliographischen Kopplung ist Null, wenn der Durchschnitt der Referenzenlisten leer ist, und erreicht sein Maximum von Eins, wenn beide Listen identisch sind (weil dann auch Durchschnitt $\mathbf{R}_i \cap \mathbf{R}_j$ und Vereinigung $\mathbf{R}_i \cup \mathbf{R}_j$ gleich sind).

Ein anderes relatives Maß der Ähnlichkeit von Mengen, das hier verwendet werden kann, ist der *Salton-Index*:[16]

$$S_{ij} = \frac{|\mathbf{R}_i \cap \mathbf{R}_j|}{\sqrt{|\mathbf{R}_i||\mathbf{R}_j|}}. \tag{3.6}$$

Hier wird die Größe des Durchschnitts der Mengen auf das geometrische Mittel der Größen beider Mengen bezogen. Auch er erreicht sein Maximum von Eins für identische Mengen und sein Minimum von Null für disjunkte.

Der Salton-Index ist immer dann dem Jaccard-Index vorzuziehen, wenn die betrachteten Mengen sehr unterschiedlich groß sein können. Beim Jaccard-Index wird dann faktisch durch die Größe der größeren Menge dividiert, beim Salton-Index nur durch die Quadratwurzel ihrer Größe.

Artikel einer Bibliographie bilden nicht nur mit den in ihnen zitierten Quellen ein bipartites Netzwerk. Auch Schlüssel- oder Titelwörter charakterisieren den Inhalt eines Artikels und können sinnvoll als zweite Knotenart verwendet werden. Im Abschnitt 3.6 (S. 35) wird das auf Salton zurückgehende Vektorraum-Modell behandelt, wo auch Terme betrachtet werden, die irgendwo im Dokument auftreten. Problematisch ist bei allen diesen Zugängen vor allem die Mehrdeutigkeit von Wörtern (Homonym-Problem) und das Auftreten verschiedener Wörter für einen Begriff (Synonym-Problem).

3.5 Kozitationsanalyse

Eine Kozitation zweier Artikel liegt dann vor, wenn beide in einem dritten zitiert werden, wie es der Graph in Abbildung 3.3 illustriert: In Artikel 1 werden die Artikel 2 und 3 zitiert. Die Kozitation kann also als das Gegenstück zur bibliographischen Kopplung zweier Artikel angesehen werden.

Auch kozitierte Artikel findet man mehrere im Netzwerk der zwölf ersten Artikel zu *N-rays* (Abb. 3.1, S. 26): Artikel 1 und 2 werden zweimal kozitiert (von 3 und 4), 8 und 9 dreimal und 10 einmal mit 8 und einmal mit 9.

Statt des Skalarprodukts der Zeilenvektoren, das wir bei der Berechnung der Matrix der bi-

[13] und die nur im Dreieck unter der Hauptdiagonalen überhaupt Elemente ungleich Null enthält

[14] vom lateinischen *ad-filiare*: als Sohn adoptieren

[15] Der schweizer Botaniker und Pflanzenphysiologe Paul Jaccard (1868–1944) definierte den Index 1901.

[16] Der in den USA wirkende Computer-Scientist Gerard Salton (1927–1995) war führend auf dem Gebiet des *information retrieval* (s. a. Wikipedia). Der Index wird im Buch von Salton und McGill (1983) verwendet, s. a. S. 128 der deutschen Ausgabe (Salton und McGill 1987). Im Abschnitt 3.6 (S. 35) wird gezeigt, wie die Autoren den Index alternativ als Cosinus des Winkels zwischen den Zeilenvektoren von Matrix A definieren.

bliographischen Kopplung B im vorigen Abschnitt bildeten, müssen wir jetzt das Skalarprodukt der Spaltenvektoren von A bilden, um die Elemente der Kozitationsmatrix C zu berechnen: $c_{ij} = \sum_k a_{ki}a_{kj}$. Da A eine binäre Matrix ist, ergibt die Summierung die Zahl der übereinstimmenden Komponenten der jeweiligen Spaltenvektoren, d. h. die Zahl der Fälle, in denen die Artikel in der gleichen Zeile bzw. Referenzenliste auftreten. Kompakt geschrieben, berechnen wir $C = A^{\mathrm{T}}A$. Der Modell-Leser bewegt sich also im Graphen zuerst gegen die Pfeilrichtung und im zweiten Schritt mit ihr. Für das Zitationsnetzwerk des betrachteten Beispiels berechnen wir Matrix C zu

$$\begin{pmatrix}
3 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
2 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 3 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 3 & 3 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.$$

Wie Matrix B ist auch C symmetrisch: $C^{\mathrm{T}} = (A^{\mathrm{T}}A)^{\mathrm{T}} = A^{\mathrm{T}}(A^{\mathrm{T}})^{\mathrm{T}} = A^{\mathrm{T}}A = C$. Die Hauptdiagonale von C enthält die Zahlen der Fälle, in denen die Artikel mit sich selbst kozitiert werden, was bei jeder Zitierung der Fall ist. Die Zahl c_{ii} ist also die Zahl aller Zitierungen von Artikel i in anderen Artikeln des Netzwerks.

Die meisten Elemente der Kozitationsmatrix C sind gleich Null. Das ist auch deswegen so, weil die inhaltlichen Beziehungen zwischen den beiden Zitationssträngen erst im ersten Überblicksartikel zu *N-rays* (Nr. 75 der Bibliographie) herausgestellt und durch Kozitation dokumentiert wurden. Kozitationsbeziehungen unterliegen also – im Gegensatz zur bibliographischen Kopplung – einem Wandel. Manch inhaltlicher Bezug wird erst später erkannt oder – umgekehrt – von späteren Autoren nicht mehr als wesentlich angesehen.

Ganz wie bei der bibliographischen Kopplung von Artikeln ist es auch bei der Kozitation sinnvoll, Artikel eines Jahrgangs mit ihren Referenzenlisten zu analysieren. Das im vorigen Abschnitt eingeführte bipartite Netzwerk von Artikeln und Quellen wird jetzt nicht daraufhin untersucht, wie die Artikel über Quellen gekoppelt werden, sondern – genau umgekehrt – wie durch Kozitationen in Artikeln die Quellen miteinander verbunden werden. Die Matrix C kann ebenfalls aus der Rechteckmatrix des bipartiten Netzwerks gemäß $C = A^{\mathrm{T}}A$ berechnet werden.

Bei der Kozitationsanalyse zweier aufeinanderfolgender Jahrgänge werden viele der im ersten Jahr hoch kozitierten Quellen auch im zweiten Jahr durch Kozitation verbunden sein. Die Kopplungsstärke wird aber variieren, neue hoch kozitierte Paare von Quellen kommen hinzu. Die Referenzenlisten eines Jahrgangs stellen quasi das Ergebnis einer Meinungsumfrage dar, bei der gefragt wird, welche Quellen aktuell als zusammenhängend angesehen werden.

Das Prinzip der Kozitation wurde zuerst von Irina Marshakova (1973) in Moskau bei einer Studie zur Laserphysik angewandt. Unabhängig von ihr propagierte es auch Henry Small (1973), Mitarbeiter am von Garfield gegründeten *Institute for Scientific Information* (ISI, Philadelphia).

In den im *Web of Science* erfassten Journalen wurde Smalls Arbeit öfter zitiert als die in russischer Sprache publizierte von Marshakova (s. Tabelle 3.1). Der anglo-amerikanische *bias* der Datenbank wirkt sich damit zweifach aus:

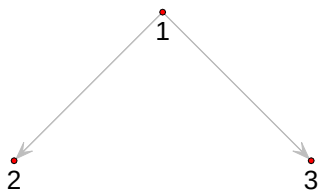


Abbildung 3.3: Kozitation: Die Artikel 2 und 3 werden beide in Artikel 1 zitiert.

Tabelle 3.1: Zeitreihe der Zitierungen und Kozitierungen der beiden grundlegenden Arbeiten zur Kozitationsanalyse: **H** bzw. **I** = Menge der Artikel, die Henry Smalls bzw. Irina Marshakovas Arbeit zitieren, J = Jaccard-Index, S = Salton-Index der Kozitation. Datenquelle: *Web of Science*, April 2008.

Jahre	$\lvert\mathbf{H}\rvert$	$\lvert\mathbf{I}\rvert$	$\lvert\mathbf{H}\cap\mathbf{I}\rvert$	J (%)	S (%)
1974–78	34	6	5	14	35
1979–83	32	10	8	24	45
1984–88	49	12	8	15	33
1989–93	65	13	10	15	34
1994–98	34	6	6	18	42
1999–03	68	10	9	13	35
2003–07	84	12	10	12	31

es fehlen Zeitschriften, in denen Marshakova zitiert wird und in der englischsprachigen Welt konnte man ihren Aufsatz nicht lesen. Die Asymmetrie kommt im Jaccard-Index der Kozitation beider Artikel zum Ausdruck, der Salton-Index gleicht die unterschiedliche Zitierhäufigkeit etwas aus (s. die beiden letzten Spalten der Tabelle 3.1 und die Gleichungen 3.5 und 3.6, S. 32).

Kozitationsanalysen können Daten für die Kartographierung der Forschung (*mapping of science*) liefern. Dazu wird das Netzwerk kozitierter Quellen aus den Referenzenlisten der Publikationen eines Jahrgangs berechnet, wobei nur Quellen einbezogen werden, welche mit ihrer Zitierungszahl einen bestimmten Schwellwert überschreiten (z. B. 5). Diese Quellen werden von Small als Symbole für Konzepte angesehen (*concept symbols*). Im so konstruierten Kozitationsnetzwerk versucht man nun, Cluster von Quellen zu bestimmen, welche enge Beziehungen zwischen Quellen des gleichen Clusters aufweisen, aber nur schwache zwischen Quellen in unterschiedlichen Clustern.[17]

Für die Ermittlung von Clustern ähnlicher Objekte, welche dieser Zielstellung möglichst gut entsprechen, sind eine Reihe von Algorithmen entwickelt worden. Wollen wir sie hier anwenden, muss zuerst entschieden werden, welches Maß für die Stärke der Kozitation zweier Quellen zu Grunde gelegt wird, die absolute Zahl von Kozitierungen oder, z. B., eines der beiden oben eingeführten relativen Maße, Jaccard- oder Salton-Index.

Die relativen Kozitationsmaße ergeben eine geringe Kopplungsstärke bei hochzitierten Quellen, welche nur selten kozitiert werden. Das ist angemessen, denn viele der zitierenden Autoren sehen ja keinen engeren Zusammenhang zwischen den zitierten Konzept-Symbolen. Am ISI wurde von Small zuerst mit dem Jaccard- und später mit dem Salton-Index gearbeitet (Small und Sweeney 1985). Marshakova (1973) benutzte kein relatives Maß, sondern berechnete die Erwartungswerte für die Kozitationszahlen bei Unabhängigkeit beider Zitiervorgänge, nämlich

$$\frac{c_{ii}}{n}\frac{c_{jj}}{n}n$$

(mit n als der Zahl aller zitierenden Aufsätze), und ließ nur die Kozitationszahlen gelten, die den jeweiligen Erwartungswert signifikant überschreiten.[18]

Im einfachsten Cluster-Verfahren wird dann ein Schwellwert der Kozitationsstärke bestimmt, welcher angibt, wie stark die Kozitation mindestens sein muss, damit die entsprechenden Links als wesentlich für das Clustering angesehen werden. Man stelle sich vor, dass mit einem Schwellwert $s = 0$ begonnen wird, durch den keine Links aus dem Kozitationsnetzwerk entfernt werden. Dann fährt man s langsam hoch, wodurch mehr und mehr Links verschwinden. Große Haufen nur schwach gekoppelter Knoten werden dabei in kleinere zerfallen, die stärker zusammenhängen. Bei einem geeigneten Schwellwert stoppt man dann.

Bei dieser von Small verwendeten Methode braucht ein Knoten nur mit einem Knoten eines Clusters verbunden zu sein, um zum Cluster zu gehören. Deswegen heißt sie *Methode des nächsten Nachbarn* oder *single-linkage clustering*. Offenbar entsprechen die so konstruierten Cluster nur unvollkommen dem oben definierten Ziel des Clustering. Es kann zum Beispiel eine längere Kette von Knoten – deren Enden nichts miteinander zu tun haben – ein Cluster bilden (*chaining*-Effekt). Bessere Verfahren sind aber auch aufwendiger und deshalb für größere Netzwerke nicht anwendbar.

Für die Analyse der Laser-Physik hat Irina Marshakova (1973) eine Clusterdefinition verwendet, die der oben formulierten Zielstellung näher kommt. Sie akzeptiert nur Cluster, bei denen die Summe der Kopplungsstärken innerhalb des Clusters größer ist als die Summe der Stärken von Kopplungen der Cluster-Knoten an Knoten außerhalb des Clusters. Ein Cluster muss also eine positive Bilanz von Innen- zu Außenbeziehungen aufweisen.[19] In Ketten sind nur wenige der internen Beziehungen realisiert (bzw. stärker als der Schwellwert). Sie werden daher nur selten Marshakovas Kriterium erfüllen, womit die Gefahr des *chaining* gebannt ist.

Um von den Clustern von Konzept-Symbolen zu einer Karte zu kommen, werden als nächstes alle Knoten je eines Clusters zu einem Punkt zusammengezogen und alle Links zwischen je zwei Clustern zu einem Link bestimmter Stärke, der eine gewisse fachliche Entfernung der Cluster entspricht. So entsteht ein übersichtliches Netzwerk von Clustern, das es zu visualisieren

[17] In der Literatur zum Web nennt man solcherart Cluster von Webpages auch *communities*.

[18] Signifikantes Überschreiten auf 5%-Niveau liegt dann vor, wenn die empirisch gefundene Kozitationszahl bei vorausgesetzter Unabhängigkeit nur in höchstens 5% der Fälle zufällig entstehen würde, wenn man genügend oft würfelt, vgl. S. 180 in dem Buch von Irina Marshakova (1988). Im Gegensatz zum Salton- und zum Jaccard-Index sind die so errechneten Zahlen von der Zahl aller zitierenden Aufsätze n abhängig.

[19] In der Literatur zum Web wird dieses Kriterium in der so genannten schwachen Definition von *communities* verwendet; die starke Definition verlangt, dass für jeden einzelnen Knoten eines Clusters die Bilanz von Innen- zu Außenbeziehungen positiv ist (Radicchi u. a. 2004).

gilt. Früher wurden dazu meist Projektionen des hochdimensionalen Netzwerks benutzt, welche die tatsächlichen Entfernungen möglichst wenig verändern (*multidimensional scaling*, MDS). Heute werden Netzwerke oft mit Verfahren des *force-directed placement* (FDP) visualisiert. Sie gehen von einer abstoßenden Kraft zwischen allen Knoten aus und einer anziehenden zwischen den über Kanten verbundenden Knoten, welche man sich als kleine Spiralfedern vorstellen kann. Realisiert man dieses Modell in drei Raumdimensionen, wird sich ein Gleichgewicht einstellen, bei dem die Federn möglichst wenig gespannt sind. Zu einer zweidimensionalen Darstellung gelangt man dann, bildlich gesprochen, indem man das Modell zwischen zwei Glasplatten zusammenquetscht.

Da die ermittelten Cluster von Knoten wiederum Knoten eines Kozitationsnetzwerkes darstellen, können mit analoger Clusterprozedur nun Cluster von Clustern ermittelt werden, und das so weiter bis alle in einem Jahrgang zitierten Quellen in einem Cluster vereinigt sind, das den in der zugrunde gelegten Datenbank indizierten Teil der Wissenschaft repräsentiert.

Gerade für Fachgebiete, in denen ein Artikel im Durchschnitt vergleichsweise nur wenige Zitierungen pro Jahr erhält, werden gerne Kozitationsanalysen von Aggregaten von Artikeln durchgeführt. Man kann danach fragen, wie oft Autoren zusammen zitiert werden, um die Struktur einer Fachgemeinschaft abzubilden. Oder man analysiert Kozitationen von Journalen. In beiden Fällen stehen jedoch die Knoten des Kozitationsnetzwerks nicht unbedingt für eine bestimmte fachliche Thematik. Nach Bradford wissen wir, dass Arbeiten zu einem Fachgebiet breit über die Literatur gestreut sind (s. Abschnitt 2.3, S. 16). Aber auch Autoren befassen sich – manchmal sogar gleichzeitig – mit mehreren Themen, welche auch verschiedenen Fachgebieten angehören können. Ein weiteres Problem bei Autor-Kozitationen ist die Tendenz zu immer mehr Kooperation, die sich in steigenden Autorzahlen pro Artikel äußert. Frühe Studien zur Autor-Kozitation beschränkten sich auf die Analyse der Erstautoren (weil nur Erstautoren im *Citation Dictionary* der CD-ROM-Edition des SCI explizit angegeben werden), was sicher nur in Gebieten mit wenig Kooperation sinnvolle Ergebnisse liefern kann. Kozitationsanalysen von Journalen basierten bisher oft auf den Links im Zitationsnetzwerk von Journalen, die man unmittelbar den *Journal Citation Reports* des SCI entnehmen kann (vgl. Abschnitt 3.2, S. 27). Dabei wird nur gezählt, wie oft zwei Journale in irgendwelchen Artikeln in einem dritten zitiert werden. Weitaus spezifischer und damit aussagekräftiger ist aber die Kozitierung von zwei Journalen in ein und demselben Artikel (und nicht nur im selben Journal, aber in möglicherweise verschiedenen Artikeln).

3.6 Exkurs: Vektorraum-Modell

Für die maschinelle Unterstützung des *information retrieval* (IR) werden Dokumente durch die in ihnen verwendeten Terme charakterisiert. Eine überschaubare und dennoch aussagekräftige Teilmenge der Terme bilden von Indexierern oder auch von den Autoren vergebene Schlüsselwörter, aber auch Titelwörter können gut für das IR verwendet werden. Im Extremfall werden alle Wörter des Dokuments berücksichtigt. Dann ist auch die Häufigkeit ihres Auftretens im Dokument von Interesse, sieht man von Stoppwörtern (*stopwords*), wie *und*, *eine* u. ä., ab. Das Auftreten von Termen in Dokumenten wird durch die Term-Dokument-Matrix A beschrieben, deren Element a_{ij} angibt, wie oft in Dokument i der Term j auftaucht.

Wissenschaftliche Zeitschriftenaufsätze werden auch durch die in ihnen zitierten Quellen charakterisiert. Der Matrix A entspricht in diesem Fall die (binäre) Affiliationsmatrix des oben eingeführten bipartiten Netzwerks von Artikeln und Quellen (s. Abschnitt 3.4, S. 31). Ganz analog beschreibt aber auch die Term-Dokument-Matrix ein bipartites Netzwerk, nämlich das von Dokumenten und Termen, nur dass hier bei Berücksichtigung der Termhäufigkeiten im Dokument das Netzwerk gewichtet ist.

Dokumente werden durch die Zeilenvektoren $\vec{a}_i^{\mathrm{T}}, i = 1, \ldots, m$ der Affiliationsmatrix A dargestellt.[20] Das legt nahe, die Ähnlichkeit zweier Dokumente mittels des Winkels zwischen ihren Dokumentvektoren zu definieren. Haben zwei Dokumente keine Terme (bzw. zitierte Quellen) gemeinsam, dann sind ihre Vektoren orthogonal zueinander; Dokumente mit gleichen (relativen) Termhäufigkeiten werden durch kollineare Vektoren (solche gleicher Richtung) repräsentiert. Als Indikator der Dokumentähnlichkeit bietet sich der Kosinus des Winkels zwischen den Dokumentvektoren an, der bekanntlich für orthogonale Vektoren Null ist und für kollineare Eins. Der Kosinus des Winkels α zwischen zwei Vektoren $\vec{u}$ und $\vec{v}$ hängt mit ihrem Skalarprodukt $\vec{u}.\vec{v}$ zusammen: $\cos\alpha = \vec{u}.\vec{v}/|\vec{u}| \cdot |\vec{v}|$. Dabei ist

[20] Wir bezeichnen hier mit $\vec{a}_i$ immer Spaltenvektoren, deswegen müssen wir transponieren, um Zeilenvektoren zu erhalten.

$|\vec{u}| = \sqrt{\sum_i u_i^2}$ die (euklidische) Länge des Vektors $\vec{u}$ und das Skalarprodukt berechnet sich als Summe der Produkte der Komponenten der Vektoren: $\vec{u}.\vec{v} = \sum_i u_i v_i$. Die Summen laufen dabei von $i = 1$ bis $i = n$, der Dimension des Vektorraums.

Im Falle des Netzwerks von Artikeln und zitierten Quellen ist die Affiliationsmatrix binär:

$$a_{ij} = \begin{cases} 1 & i \to j \\ 0 & \text{sonst} \end{cases} .$$

Das Skalarprodukt zweier Zeilenvektoren von A ist dann gleich der Zahl der Quellen, die von beiden Artikeln zitiert werden, und die euklidische Länge der binäre Zeilenvektoren ist gleich der Wurzel aus der Zahl ihrer Komponenten, weil $a_{ij}^2 = a_{ij}$. Damit wird klar, dass der in Abschnitt 3.4 zur bibliographischen Kopplung (s. Gleichung 3.6) mengentheoretisch eingeführte Salton-Index für binäre Dokumentvektoren mit dem hier definierten Kosinus-Index zusammenfällt.

Im Vektorraum-Modell des IR werden aber nicht nur die Ähnlichkeiten von Dokumenten über die in ihnen verwendeten Terme ermittelt, sondern auch die Zusammenhänge zwischen Termen über ihre gemeinsame Verwendung in den Dokumenten. Ersteres entspricht bei Artikeln und Quellen der bibliographischen Kopplung, letzteres der Kozitation von Quellen (Abschnitt 3.5, S. 32).

Bipartite Netzwerke von Artikeln und zitierten Quellen haben gegenüber Term-Dokument-Netzwerken den Vorteil, dass sie unabhängig von der Sprache des Dokuments sind.

3.7 Koautorschafts-netzwerke

Koautorschaft gilt als Indikator für Kooperation. Wenn zwei oder mehr Autoren gemeinsam eine Publikation von Forschungsergebnissen verantworten, dann sollten sie im Forschungsprozess, der zu diesen Ergebnissen führte, in irgendeiner Weise zusammengearbeitet haben. Es ist also angebracht, sich auf einen Begriff von Kooperation zu verständigen, bevor die Struktur von Koautorschaftsnetzwerken analysiert und interpretiert wird.[21]

Wenn bei der Analyse der Wissensproduktion nicht von deren immanent sozialem Charakter ausgegangen wird, bleiben die Ergebnisse notwendig mager. Obwohl Ideen im einzelnen Forscherhirn entstehen, so ist dieser Vorgang unmöglich ohne Anregung durch andere und folgenlos ohne Adressaten. Erst nur über Lehrer, dann auch über Literatur kam älteres Wissen auf die Gelehrten jeder neuen Generation; dann gab es von ihnen so viele, dass sie untereinander in größerem Maße kommunizieren konnten. Das neunzehnte Jahrhundert brachte die Gruppe von Forschern, welche sich einem Projekt widmen, das zwanzigste die Zusammenarbeit von Instituten und Staaten in der Forschung. Will man für quantitative Untersuchungen Zusammenarbeit in der Forschung begrifflich scharf von individueller Arbeit abgrenzen, so stößt man auf Schwierigkeiten, weil Wissensproduktion eines Einzelnen nicht ohne die anderer Menschen möglich ist.

Katz und Martin (1997, S. 7) kommen beim Versuch der Definition von Forschungskooperation zu dem unbefriedigenden Ergebnis, dass wenn alle, die zu einem Forschungsresultat direkt oder indirekt beigetragen haben, als Kooperationspartner aufgeführt werden würden, die Liste kein Ende hätte, und im anderen Extrem die Liste strenggenommen oft ganz leer bleiben müsste, wenn sie nur die enthielte, die das Projekt vollständig überblicken und verantworten können.

Grit Laudel (1999, S. 32–40) entrinnt dieser Unbestimmtheit, indem sie Forschungskooperation nicht in Bezug auf das Ergebnis, sondern auf das Handeln der Forscher fasst, welches nach ihrer Auffassung notwendig aufeinander bezogen sein muss, um es kooperativ nennen zu können. In solchen Fällen, wo die Partner in einem gemeinsamen Arbeitsprozess kreativ zum Fortgang des Projektes beitragen und ihre Beiträge gemeinsam in das Ergebnis integrieren, spricht sie von *arbeitsteiliger* Kooperation (S. 40). Solcherart Forschungshandeln als kooperativ zu bezeichnen, ist unproblematisch. Schwieriger wird es bei den Fällen, die sie der *unterstützenden* Zusammenarbeit zurechnet, und bei denen der unterstützende Partner Routine-Dienstleistungen, vorhandene Geräte oder vorhandenes Wissen zur Verfügung stellt. Vorhandenes Wissen kann auch der Literatur entnommen werden und serienmäßig hergestellte Forschungsgeräte können auf dem Markt erworben werden. Hier von Kooperation zu sprechen, ist nach der Laudel'schen Definition jedoch ausgeschlossen, weil Literatur und Geräte unspezifisch erzeugt wurden, nicht auf das Projekt bezogen. Analog kann bei den Dienstleistungen abgegrenzt werden.

[21] Die folgenden Absätze zur Definition von Forschungskooperation hat der Autor bereits 2002 in einer für das Bundesministerium für Forschung und Technologie verfassten Expertise verwendet (vgl. Online-Publikation auf `http://www.sciencepolicystudies.de`).

Arbeitsteilige Kooperation, so sie erfolgreich war, wird sich fast immer im gemeinsamen Publizieren der Forschungsresultate äußern, bei der unterstützenden Zusammenarbeit sind die Gepflogenheiten nicht einheitlich. Routinearbeit bei der Messung und Verarbeitung der Daten, Programmierarbeit, das Beisteuern von Messproben oder -geräten und nicht zuletzt die informelle Kommunikation von Wissen sind Formen der Zusammenarbeit, die in der Danksagung berücksichtigt aber auch mit Koautorschaft belohnt werden können. Bibliometrisch wird mit der Analyse der Koautorschaften also vor allem die arbeitsteilige Kooperation erfasst, wenn auch – ein wesentlich aufwendigeres Unterfangen – Danksagungen ebenfalls schon untersucht wurden. Beim Vergleich der Koautorschaftsdaten verschiedener Länder, Fachgebiete oder Zeiträume muss beachtet werden, dass die unterstützende Kooperation sich möglicherweise in jeweils unterschiedlichem Maße in Koautorschaft äußert.

Nicht jede Zusammenarbeit im oben definierten Sinne äußert sich also in Koautorschaft. Andererseits kann auch eine gewisse Tendenz zu unbegründetem Vergeben (oder Erzwingen) von Koautorschaften nicht ganz negiert werden. Trotzdem wird die gemeinsame Verantwortung eines Aufsatzes in einer renommierten Fachzeitschrift nur selten ohne irgendeine Art von Zusammenarbeit möglich sein. Zumindest kennen sich Koautoren.[22] Diese realistische Annahme macht die Analyse von Koautorschaftsnetzwerken interessant.

Sie werden meist als Netzwerke von Autoren eingeführt. In der einfachsten Form ist das Koautorschaftsnetzwerk ungewichtet. Ein Link zwischen zwei Autoren liegt vor, wenn sie in mindestens einer Publikation der untersuchten Bibliographie beide als Autoren auftreten. Eine Wichtung des Links mit der Zahl der Artikel, bei denen beide als Autoren auftreten, liegt nahe. Aber auch diese Modellierung benutzt nur einen Teil der Information, die über das Zusammenwirken von Autoren aus einer Bibliographie entnommen werden kann. Es bleibt unberücksichtigt, ob die Beziehungen jeweils rein bilateral sind oder ob Forscher auch in größeren Gruppen zusammenarbeiten. Wenn zum Beispiel drei Autoren eine Arbeit zusammen verfassen, sind zwischen ihnen im Netzwerk genauso drei Links mit dem Gewicht 1 vorhanden, wie wenn je zwei von ihnen jeweils immer einen Aufsatz zusammen publizieren (insgesamt also drei Aufsätze).

[22] Das gilt sicher nicht uneingeschränkt für Artikel mit hundert und mehr Autoren.

Die volle Information wird genutzt, wenn Koautorschaft in einem bipartiten Netzwerk von Autoren und Artikeln dargestellt wird. Element a_{ij} der Affiliationsmatrix A ist gleich Eins, wenn Autor i in der Autorenliste von Artikel j auftritt und sonst gleich Null.

Die meisten Untersuchungen beschränken sich aber bislang auf Koautorschaftsnetzwerke, in denen nur Autoren durch Knoten repräsentiert sind und nicht die Publikationen. Die Adjazenzmatrix B eines solchen Netzwerks ist aus der Affiliationsmatrix A über $B = AA^{\mathrm{T}}$ berechenbar, was man sofort einsieht, wenn man die Überlegungen zum Netzwerk bibliographisch gekoppelter Artikel auf das Autoren-Artikel-Netzwerk überträgt (s. Abschnitt 3.4, S. 31). Ein Koautor ist jemand, den man erreicht, indem man im bipartiten Netzwerk einen Schritt zu einer Publikation macht (A^{T}) und dann wieder einen zurück zu einem Autor (A). Die Koautorschaftszahlen zweier Autoren erhält man durch das Skalarprodukt der (binären) Zeilenvektoren von A. Das Diagonalelement b_{ii} von Matrix B ist gleich der Zahl der Publikationen, an denen Autor i beteiligt ist.

Wie sind nun Koautorschaftsnetzwerke strukturiert? Zum einen findet man oft, dass die große Mehrheit (über 80%) von Autoren eines Fachgebiets in einer Komponente des Netzwerks (genannt Hauptkomponente, *main component*) versammelt ist. Die restlichen Autoren bilden dagegen oft nur kleine Gruppen von über Koautorschaftslinks (wenigstens indirekt) miteinander verbundenen Forschern. Alle zwischen Kooperationspartnern auftretenden Distanzen – fachliche, institutionelle, geographische, sprachliche, kulturelle, politische – verhindern nicht das Entstehen eines großen zusammenhängenden Netzwerks von kooperierenden Wissenschaftlern.

Ebenso bemerkenswert sind die im Vergleich zur Größe geringen Entfernungen zwischen Autoren in den Hauptkomponenten von Koautorschaftsnetzwerken. Im Koautorschaftsnetzwerk von über einer Million biomedizinischer Autoren, die in den fünf Jahren 1995–99 zusammen über zwei Millionen (in Medline nachgewiesene) Artikel publizierten, fand der statistische Physiker Mark E. J. Newman (2001a) über 90% der Autoren in der Hauptkomponente.[23] Die zweitgrößte Komponente enthält nur 49 Au-

[23] Da Autoren in bibliographischen Datenbanken nicht immer eindeutig identifizierbar sind, schwanken Autorenzahlen mit der Methode ihrer Identifizierung. Newman fand bei Berücksichtigung des vollen Namens mehr als 1.5 Millionen unterschiedliche Autoren, wenn er neben dem Familiennamen nur die Initiale des ersten Vornamens berücksichtigt, schmilzt die Zahl auf knapp 1.1 Millionen zusammen.

toren. Die maximale Entfernung zwischen zwei Autoren in der Hauptkomponente (auch Durchmesser des Netzwerks genannt) beträgt nur 24 Schritte (*hops*), d. h. auf dem kürzesten Weg (*shortest path*) zwischen zwei beliebigen Knoten liegen höchstens 23 andere Knoten. Die mittlere Länge aller kürzesten Pfade zwischen Knoten der Hauptkomponente ist kleiner als fünf Schritte (Newman 2001b):

> This "small world" effect, first described by Milgram,[24] is, like the existence of a giant component,[25] probably a good sign for science; it shows that scientific information—discoveries, experimental results, theories—will not have far to travel through the network of scientific acquaintance to reach the ears of those who can benefit by them. (Newman 2001b, S. 3)

Diese Einschätzung Newmans beschränkt sich auf die informelle Kommunikation von Resultaten zwischen Forschern, die ja aber oft der formellen Publikation vorausgeht. Das berühmte, vom Sozialpsychologen Stanley Milgram (1967) unternommene Experiment ergab für das Bekanntschaftsnetzwerk in den USA eine mittlere Distanz von sechs Schritten.[26] Newman erklärt den *small-world*-Effekt an sich selber: Er hat 26 Koautoren, die wiederum mit insgesamt 623 anderen Forschern zusammen Publikationen verfasst haben:

> The "radius" of the whole network around me is reached when the number of neighbors within that radius equals the number of scientists in the giant component of the network, and if the increase in numbers of neighbors with distance continues at the impressive rate [...], it will not take many steps to reach this point. (Newman 2001b, S. 3)

Die Zahl von Koautoren eines Autors ist ein Maß seiner Vernetzung. Sie ist gleich der Zahl seiner Links (genannt Grad bzw. *degree* des Knotens) im Koautorschaftsnetzwerk. Im Fachgebiet randständige Autoren haben nur wenige Kooperationspartner. In der Netzwerkanalyse dient deshalb der Grad eines Knotens auch als Maß für seine Zentralität (*centrality*). Oft ist die Koautorenzahl schief verteilt: wenige Autoren haben viele Koautoren, viele nur wenige (Newman 2001a, S. 5).

Ein anderes Zentralitätsmaß ist die *betweennes* eines Knotens i, die definiert ist als die Gesamtzahl der kürzesten Pfade zwischen beliebigen Knotenpaaren, welche durch Knoten i verlaufen. Man kann sich vorstellen, dass Knoten mit hoher *betweennes* für kurze Distanzen im Netzwerk und auch für das Entstehen einer großen Hauptkomponente verantwortlich sind. Auch bei der *betweennes centrality* setzen sich in Koautorschaftsnetzwerken einige Spitzenreiter klar von den übrigen Autoren ab (Newman 2001b, S. 2).

[24] Milgram (1967)

[25] Newman (2001a)

[26] Seine Probanden hatten die Aufgabe, über Bekannte einen Brief näher an eine bestimmte ihnen unbekannte Person heranzubringen. Die Briefe, die bei den Zielpersonen ankamen, waren durchschnittlich von sechs Personen (inklusive der Probanden) weitergesandt worden.

Kapitel 4

Bibliometrische Modelle

4.1 Der Matthäus-Effekt

Der Evangelist Matthäus erzählt das Gleichnis Jesu von den anvertrauten Zentnern nach: Drei Knechte bekommen von ihrem Herrn fünf, zwei und einen Zentner (Talente) Silber zur Verwahrung während seiner Reise. Die ersten beiden wuchern und verdoppeln dadurch ihre Zentner, während der dritte seinen einen furchtsam vergräbt. Nach der Rückkehr des Herrn schenkt er den ersten beiden das ganze Silber, der furchtsame bekommt nichts, denn der gestrenge Herr urteilt (Matthäus, 25, 28–30):

> Darum nehmt von ihm den Zentner und gebt es dem, der zehn Zentner hat. Denn wer da hat, dem wird gegeben werden, und er wird die Fülle haben; wer aber nicht hat, dem wird auch, was er hat, genommen werden. Denn wer da hat, dem wird gegeben, daß er die Fülle habe; wer aber nicht hat, von dem wird auch das genommen was er hat. Und den unnützen Knecht werft hinaus in die Finsternis; da wird sein Heulen und Zähneklappen.[1]

Das sprichwörtliche Prinzip des "Wer hat, dem wird gegeben" gilt nach Robert K. Merton (1910–2003), dem bekannten US-amerikanischen Soziologen auch in der Wissenschaft. Er prägte den Ausdruck Matthäus-Effekt (*Matthew effect*), der mittlerweile nicht nur innerhalb der Wissenschaftssoziologie gebräuchlich geworden ist (Merton 1968).

Es leuchtet unmittelbar ein, dass die großen Ungleichheiten in Gesellschaft, Wirtschaft und Wissenschaft etwas mit einem solchen Effekt zu tun haben. Es scheint so zu sein, dass Erfolg neuen Erfolg bringt (*success breeds success*), dass Vorteile kumulieren (*principle of cumulative advantage*) und dass die Reichen leichter reicher werden (*the rich get richer*).[2] Ein Guthaben um 10% zu steigern, ist immer gleich schwer, egal ob es sich um 100 Euro oder um eine Million handelt.

Ich kann hier nicht im Einzelnen diskutieren, welche konkreten Mechanismen den Matthäus-Effekt in der Wissenschaft hervorrufen; nur ein paar Sätze zu diesem wissenschaftssoziologischem Thema möchte ich einrücken. Wissenschaftler verkaufen das von ihnen produzierte Wissen nicht, sondern streben nach Reputation, indem sie es öffentlich machen. Reputation befähigt sie, gut dotierte Stellen zu erlangen. Voraussetzung für Reputation ist die Aufmerksamkeit der Fachkollegen – bekanntlich ein rares Gut. Sie wird – wie in Kunst, Sport und Politik – vor allem denen gegeben, die schon viel davon bekommen haben. Reputation führt aber auch zu Forschungsmitteln und damit zu neuen Chancen für wissenschaftlichen Erfolg. All das – und noch einiges mehr – bewirkt eine selbstverstärkende Rückkopplung in den Karrieren von Forschern. Ähnliche Betrachtungen kann man über den Matthäus-Effekt bei wissenschaftlichen Institutionen und Zeitschriften anstellen.

Der Matthäus-Effekt wird aber nicht nur für eine qualitative Erklärung von Ungleichheiten verwendet, sondern auch in mathematische Modelle eingebaut, die schiefe Verteilungen zum Resultat haben, wie sie gerade für die Wissenschaft charakteristisch sind (vgl. Kapitel 2, S. 13ff.). Mathematische Modelle sind vereinfachende quantitative Beschreibungen von Phänomenen, die helfen, die Phänomene zu erklären, d. h. sie auf das Wirken plausibler Prinzipien zurückzuführen. Es geht also z. B. darum, die Lotka-Verteilung auf den Matthäus-Effekt zurückzuführen. Der erste, der ein derartiges Modell für die Erklärung des Potenzgesetzes der wissenschaftlichen Produktivität heranzog, war der einflussreiche US-amerikanische Sozialwissenschaftler und Nobelpreisträger für Wirtschaftswissenschaften 1978 Herbert A. Simon (1916–2001).[3] Er stützte sich dabei auf Ar-

[1] `http://www.bibel-online.net/buch/40.matthaeus/25.html`

[2] Im Deutschen heißt es drastischer: "Der Teufel scheißt immer auf den größten Haufen."

[3] Quelle: Wikipedia

beiten des bekannten schottischen Statistikers George Udny Yule (1871–1956) in den 1920-er Jahren, in denen dieser ein Modell für die Verteilung biologischer Arten auf Familien konstruiert hatte, welche ebenfalls schief ist (viele Familien haben wenige Arten, wenige viele) und ebenfalls einem Potenzgesetz folgt (Simon 1955).

4.2 Der Yule-Prozess

Der Yule-Prozess kann mit Bezug auf Autoren und Artikel einer Bibliographie im einfachsten Fall so beschrieben werden: Zu Beginn gibt es einen Autor mit einer Publikation in der Bibliographie. In jeder Runde des Prozesses werden der Bibliographie zwei Artikel hinzugefügt, und zwar so, dass einer von einem neuen Autor publiziert wird und einer von einem Autor, der bereits in der Bibliographie vertreten ist. Der zweite Artikel wird unter den bisherigen Autoren verlost, wobei jeder Autor für jeden seiner bisherigen Artikel ein Los erhält. Ein Autor mit bisher zehn Artikeln hat damit eine zehnfach größere Chance, einen weiteren zu publizieren, als ein Autor mit nur einer Publikation. Auf diese Weise wird im Modell der Matthäus-Effekt hervorgerufen: Wer hat, dem wird gegeben.

Interessant für uns ist nun, zu welcher Verteilung von Artikeln auf Autoren der Yule-Prozess führt, wenn genügend Zeit verstrichen ist, so dass sich eine Verteilung stabilisieren kann, d. h. dass die Anteile der Autoren mit einem Artikel, mit zwei, drei usw. Artikeln nicht mehr stark schwanken. Dabei ist es noch unklar, ob der Prozess überhaupt zu einer stabilen Verteilung konvergiert. Möglich wäre auch, dass die Anteile sich ständig weiter verschieben. Yule leitete eine finale Verteilung ab, die von Simon (1955) in einer modernen Sprache neu hergeleitet und Yule-Verteilung getauft wurde.

Im allgemeinen Fall werden pro Runde m Artikel verlost, so dass die Bibliographie jeweils um $m + 1$ Artikel wächst. Da in jeder Runde ein Autor hinzukommt, sind die Autoren zu jedem Zeitpunkt im Mittel mit $m+1$ Publikationen in der Bibliographie vertreten. Als zeitunabhängige Verteilung der Anteile $F(j)$ von Autoren mit j Artikeln erhält man die Yule-Verteilung

$$F(j) = (1 + 1/m) \cdot \mathrm{B}(j, 2 + 1/m). \qquad (4.1)$$

Hierbei bezeichnet B die Betafunktion, die mittels der Gammafunktion definiert wird:

$$\mathrm{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}. \qquad (4.2)$$

Die Gammafunktion ist durch ein Integral definiert und vereinfacht sich für positive ganzzahlige Argumente $n = 1, 2, 3, \ldots$ auf die einfache Fakultätsfunktion: $\Gamma(n) = (n - 1)! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot (n - 1)$.[4]

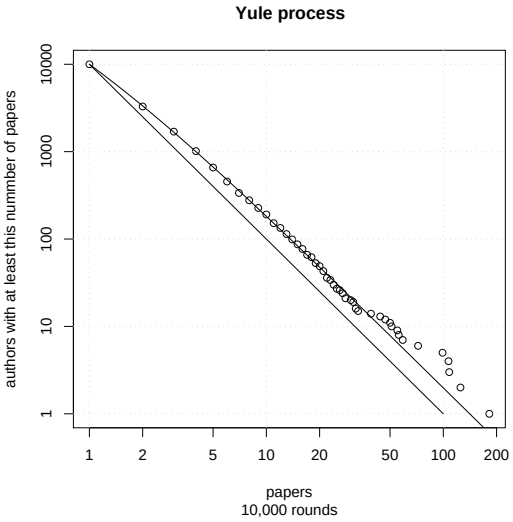


Abbildung 4.1: Zipf-Plot einer Simulation des Yule-Prozesses mit $m = 1$ nach 10.000 Runden

Jetzt können wir die Verteilung bestimmen, die bei der oben erwähnten einfachsten Variante des Yule-Prozesses nach genügend langer Zeit entsteht. Hier wird nur ein Artikel pro Runde verlost, d. h. es gilt $m = 1$ und wir erhalten:

$$\begin{aligned} F(j) &= 2 \cdot \mathrm{B}(j, 3) \\ &= 2 \cdot \frac{(j-1)!2!}{(j+2)!} \\ &= 4 \cdot \frac{1 \cdot 2 \cdot 3 \cdot \ldots \cdot (j-1)}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot (j-1)j(j+1)(j+2)} \\ &= \frac{4}{j(j+1)(j+2)}. \end{aligned}$$

Diese Funktion ist zwar keine Potenzfunktion, aber sie nähert sich der Potenzfunktion $4/j^3$ immer mehr an je größer j wird. Für beliebiges m erhält man bei $j \to \infty$ ein Potenzgesetz mit $\alpha = 2 + 1/m$.

Wir lassen jetzt den Yule-Prozess mit einem verlosten Artikel pro Runde ($m = 1$) auf dem Rechner ablaufen, wozu ein Programm von wenigen Zeilen ausreichend ist. Man nennt so etwas Computersimulation, ein Verfahren, das auch zum Testen von neuen Modellen verwendet werden kann, für die noch keine Grenzverteilung mathematisch abgeleitet worden ist. Schon nach einigen Dutzend Runden stabilisiert sich die gemessene Verteilung $F(j)$. Die Abbil-

[4] Eine gut nachvollziehbare Ableitung der Yule-Verteilung findet man bei Newman (2005).

dung 4.1 zeigt die kumulierte Verteilungsfunktion nach zehntausend Runden, d. h. für 10001 Autoren mit 20001 Artikeln. Zum Vergleich ist auch für die theoretische Yule-Verteilung $F(j) = 4/j(j+1)(j+2)$ die zugehörige kumulierte Verteilungsfunktion $F_{\mathrm{cum}}(j)$ dargestellt. Sie ergibt sich zu

$$\begin{aligned} F_{\mathrm{cum}}(j) &= \sum_{k=j}^{\infty} F(k) \\ &= \sum_{k=j}^{\infty} \frac{4}{k(k+1)(k+2)} \\ &= 2\sum_{k=j}^{\infty} \frac{k+2-k}{k(k+1)(k+2)} \\ &= 2\sum_{k=j}^{\infty} \frac{1}{k(k+1)} - \\ &\quad -2\sum_{k=j}^{\infty} \frac{1}{(k+1)(k+2)} \\ &= 2/j(j+1). \end{aligned}$$

Wir erwarten also für große j einen Abfall der kumulierten Funktion mit $1/j^2$. Um diese Erwartung überprüfen zu können, ist in der Abbildung 4.1 die Funktion $1/j^2$ eingezeichnet, die wegen der doppelt-logarithmischen Darstellung als Gerade sichtbar wird. Schon für Werte ab $j = 5$ ist die Neigung der Geraden von der von $F_{\mathrm{cum}}(j)$ nicht zu unterscheiden.

Der Yule-Prozess ergibt also eine durch eine Funktion $F(j)$ beschriebene Verteilung, die für große Werte von j nach einem Potenzgesetz abfällt, das in Gesellschaft, Wirtschaft und Wissenschaft oft angetroffen wird. Man kann in diesen Fällen vermuten, dass dem ein Matthäus-Effekt zu Grunde liegt, auch wenn die empirisch ermittelten Verteilungen für kleine Werte von j nicht vollkommen mit einer Yule-Verteilung übereinstimmen. Man kann nur vermuten, weil prinzipiell auch Modelle ohne Matthäus-Effekt *power-law*-Verteilungen ergeben können.

Die in der Abbildung 4.1 deutlich sichtbaren Abweichungen der Positionen der Spitzenreiter von der jeweils erwarteten Anzahl publizierter Artikel bildet sich in zufälliger Weise schon nach einigen Runden heraus. Wegen des Matthäus-Effekts sind die Spitzenreiter dann nicht mehr einholbar.

Es ist naheliegend, einen Matthäus-Effekt auch für die schiefen Verteilungen von Zitationen auf Artikel verantwortlich zu machen. Es ist aus verschiedenen Gründen wahrscheinlicher, dass ein oft zitierter Artikel eine weitere Zitierung erhält, als ein bisher wenig zitierter. Der Yule-Prozess wurde zuerst von Derek de Solla Price (1976) mit Zitationsverteilungen in Zusammenhang gebracht. Er kann für Zitationsnetzwerke von Artikeln so beschrieben werden (Newman 2003, S. 30–31): Am Anfang gibt es eine Reihe von Artikeln, die keine Referenzen besitzen (bzw. nur Quellen außerhalb des Netzwerkes zitieren). Dann wird in jeder Runde ein Artikel hinzugefügt, welcher im Mittel m Referenzen besitzt, bzw. m Artikel im Netzwerk zitiert. Die neuen Referenzen werden unter den vorhandenen Artikeln verlost, wobei jeder Artikel so viele Lose bekommt, wie er bisher bereits Zitationen erhalten hat. Da jeder Artikel bei Erscheinen noch nicht zitiert sein kann, würde er nie an der Verlosung teilnehmen, wenn ihm nicht ein Anfangsbonus zugeteilt würde. Price nimmt der Einfachheit halber an, der Bonus würde 1 betragen und interpretiert das so, dass die Publikation selber quasi die erste Zitierung sei. Als Ergebnis erhält man für die Verteilung der Zitationen auf Artikel die Yule-Verteilung, wie sie in Gleichung 4.1 gegeben ist.

Später zog dann das Modell von Barabasi und Albert (1999) für die Linkstruktur im Web viel Aufmerksamkeit auf sich. Auch hier finden wir schiefe Verteilungen – viele Webpages haben wenig Inlinks, wenige viel – und auch hier fällt die die Verteilung beschreibende Funktion für große Inlink-Zahlen nach einem Potenzgesetz ab. Barabasi und Albert führen dies in ihrem Modell auf das so genannte *preferential attachment* zurück, d. h. auf die Tendenz, dass Webpages, auf die bereits von vielen andere Webpages ein Link gesetzt wurde, eine größere Chance haben, von einer neuen Webpage einen weiteren Link zu bekommen. Das ist nur eine neue Formulierung des Matthäuseffekts, bzw. des *principle of cumulative advantage*. Von Mark E. J. Newman (2003, S. 32) wurde darauf hingewiesen, dass das Barabasi-Albert-Modell des Web nur eine Variante des von Price auf das Zitationsnetzwerk angewandten Yule-Prozesses darstellt, und zwar eine Variante mit m Bonuspunkten für jede neue Webpage, was auf eine Yule-Verteilung $F(j) = 2 \cdot \mathrm{B}(j, 3)$ führt (die wir auch oben als Ergebnis der Simulation mit $m = 1$ diskutiert haben).

Price war nicht nur der erste, der den Yule-Prozess für die Erklärung von Netzwerkeigenschaften heranzog, er hat im gleichen Aufsatz auch ein weiteres Modell entworfen, in dem das Prinzip des kumulierenden Vorteils realisiert ist, nämlich sein Urnenmodell.

4.3 Das Urnenmodell von Price

In der Wahrscheinlichkeitstheorie dienen Urnenmodelle schon länger zur Veranschaulichung von Zufallsprozessen. Das von Derek J. de Solla Price (1976) vorgeschlagene Modell kann so beschrieben werden: Man stelle sich einen Jahrgang von N Absolventen vor, die alle ihre Abschlussarbeit publiziert haben. Jeder Absolvent bekommt nun eine Urne, die eine rote und eine schwarze Kugel enthält. Alle N versuchen ihr Glück und ziehen eine Kugel. Wer eine schwarze zieht, hat Pech und ist raus aus dem Spiel, d. h. ihm gelingt keine weitere Publikation. Die anderen publizieren einen weiteren Aufsatz, spielen weiter, legen ihre rote Kugel zurück in die Urne und bekommen noch eine weitere rote Kugel hinzu, so dass ihre Chance, in der nächsten Runde wieder eine rote zu ziehen, erhöht wird. Diese Realisierung des Prinzips des kumulierenden Vorteils wiederholt sich jede Runde, bis auch der letzte Autor eine schwarze Kugel gezogen hat. Autoren mit einer Publikation haben die Chance 1/2 weiter zu kommen, mit zwei Publikationen 2/3 und allgemein ist die Wahrscheinlichkeit für eine weitere Publikation $j/(j+1)$, wenn man bereits j Artikel publiziert hat. Die Chance für mindestens $j+1$ Artikel ist also

$$F_{\mathrm{cum}}(j+1) = \frac{1}{2}\cdot\frac{2}{3}\cdot\frac{3}{4}\cdot\ldots\cdot\frac{j-1}{j}\cdot\frac{j}{j+1} = \frac{1}{j+1}.$$

Der zu erwartende Anteil $F(j)$ von Autoren mit genau j Publikationen ist dann der Anteil von Autoren mit mindestens j vermindert um den Anteil mit mindestens $j+1$ Publikationen:[5]

$$\begin{aligned} F(j) &= F_{\mathrm{cum}}(j) - F_{\mathrm{cum}}(j+1) \\ &= \frac{1}{j} - \frac{1}{j+1} = \frac{1}{j(j+1)}. \end{aligned}$$

Wir erhalten somit für das Urnenmodell von Price als Resultat die Yule-Verteilung (Gleichung 4.2) im Grenzwert für $m \to \infty$:

$$\lim_{m\to\infty}(1+1/m)\cdot \mathrm{B}(j, 2+1/m)$$

$$= B(j,2) = \frac{(j-1)!}{(j+1)!} = \frac{1}{j(j+1)}.$$

Da aber für die Yule-Verteilung die mittlere Zahl von Publikationen pro Autor gerade $m+1$ ist, können wir für das Price'sche Urnenmodell keinen Erwartungswert für die Publikationszahl pro Autor berechnen. In jeder Realisierung des Modells wird die mittlere Publikationszahl endlich sein, aber für ihren Wert können wir keine Voraussage machen.

Für große j verschwindet $F(j)$ mit $1/j^2$. Auch für dieses Potenzgesetz errechnet man eine unendliche Zahl von Publikationen:

$$\sum_{j=1}^{\infty} j\frac{1}{j^2} = \sum_{j=1}^{\infty}\frac{1}{j} = \infty.$$

In den Fällen, wo das Lotka-Gesetz mit dem Exponenten $\alpha = 2$ auch für große Publikationszahlen j gilt, ist also die mittlere Publikationszahl eine nicht voraussagbare Größe. Das gilt genauso für $\alpha < 2$. Für $\alpha > 2$ kann man einen endlichen Wert berechnen, dessen Standardabweichung aber größer als der Wert selber ist. Das ist Ausdruck der Tatsache, dass für schiefe Verteilungen das arithmetische Mittel wenig Aussagekraft besitzt. Es ist stark von Ausreißern abhängig (vgl. Abschnitt 5.2, S. 48).

4.4 Exkurs: Gesetz von Gibrat

Auch das von dem französischen Ökonomen Robert Gibrat (1904–1980) in den dreißiger Jahren vorgeschlagene Gesetz des proportionalen Wachstums von Firmen ist Ausdruck des Matthäus-Effekts: Eine große Firma vermag genauso leicht oder genauso schwer, um z. B. 10% zu wachsen wie eine kleine. Wir realisieren ein Modell, das diesem Gesetz folgt, mittels eines kleinen Computer-Programms. Im ersten Jahr haben $N = 100\,000$ Firmen alle die gleiche Größe, die wir mit $s = 1$ ansetzen. Wie stark jede einzelne Firma wächst, ist dem Zufall überlassen. Wir ziehen als jährlichen Wachstumsfaktor jeweils eine zwischen 0.55 und 1.55 – der Einfachheit halber – gleichverteilte Zufallszahl. Die gleichverteilten Wachstumsfaktoren sind nicht realistisch, aber das Ergebnis der Simulation, die Größenverteilung der Firmen ist nach hinreichend vielen Jahren unabhängig von der Verteilung der Wachstumsfaktoren und auch von der Anfangsverteilung der Firmengrößen. Schon nach neun Jahren erhalten wir eine Größenverteilung, die sich von der korrespondierenden Lognormalverteilung wenig unterscheidet (Abbildung 4.2). Wegen der logarithmischen Skalierung der x-Achse wird die Lognormalverteilung als Gaußsche Glockenkurve angezeigt.

Eine solche Lognormalverteilung liegt für Firmengrößen empirisch oft vor. Wir sind ihr

[5] Wir erwarten $N \cdot F_{\mathrm{cum}}(j)$ Autoren, die bis zur Runde mit j Artikeln gekommen sind. Diese werden in der nächsten Runde um die $N \cdot F_{\mathrm{cum}}(j+1)$ vermindert, die eine Runde weiterkommen.

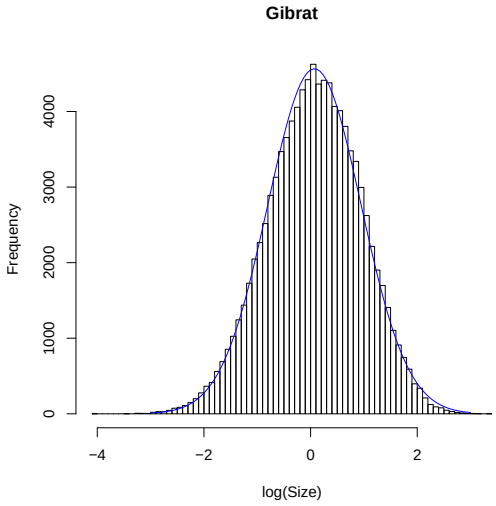


Abbildung 4.2: Histogramm einer Simulation des Gibrat-Gesetzes mit 100 000 Firmen nach neun Jahren und Wachstumsraten, die zwischen 0.55 und 1.55 gleichverteilt sind; mittleres Wachstum in neun Jahren: 7.5%; Kurve: Lognormalverteilung mit $\mu = .072$ und $\sigma = .874$ (Sample-Werte)

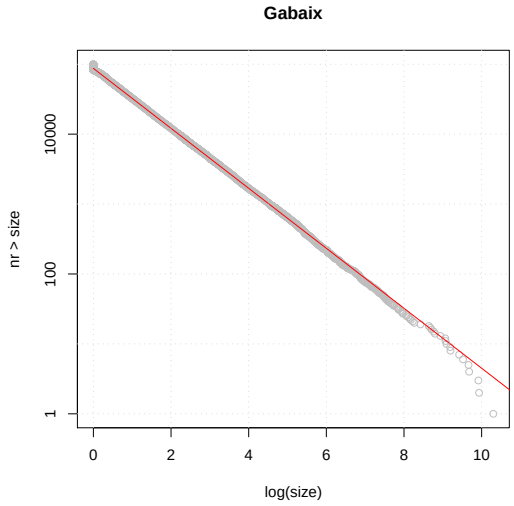


Abbildung 4.3: Zipf-Plot einer Simulation des Gibrat-Gesetzes mit minimaler Größe 1 nach Gabaix: 100 000 Firmen nach 500 Runden und Wachstumsraten, die zwischen 0.55 und 1.55 gleichverteilt sind; die Gerade ist durch lineare Regression der 80 000 größten Firmen gewonnen und entspricht einer fallenden Potenzfunktion mit Exponent $\gamma = 0.99(4)$ (die Zahl in Klammern gibt an, wie stark die letzte Stelle innerhalb der Standardabweichung streut); $R^2 = .9999$.

bei der Produktivität von Autoren, gemessen mit fraktionaler Zählweise, begegnet (s. Abschnitt 2.4, S. 18). Die Lognormalverteilung ist – wie die Yule-Verteilung – schief: Nur wenige der Akteure erreichen eine hohe Produktivität. Der Unterschied zwischen beiden liegt einerseits darin, dass bei Yule-Verteilungen keine beliebig kleinen Werte auftreten können und dass andererseits für große Werte die Häufigkeiten bei Lognormalverteilungen schneller gegen Null gehen als bei den Potenzfunktionen, die das asymptotische Verhalten der Yule-Verteilungen beschreiben.

Gauß'sche Normalverteilungen entstehen, wenn die zufallsverteilten Zuwächse nicht multiplikativ, sondern additiv wirken. Da durch die Logarithmierung die Multiplikation zur Addition wird, folgt unmittelbar, dass bei proportionalem Wachstum nach Gibrat's Gesetz eine Normalverteilung der Logarithmen, d. h. eine Lognormalverteilung zu erwarten ist.

Die von ihm gefundene Lognormalverteilung fraktionaler Produktivität von Autoren wurde von Shockley (1957) nicht mit dem Gibrat'schen Wachstumsgesetz erklärt, sondern mit einem multiplikativen Wirken von den vielen für den Erfolg nötigen subjektiven und objektiven Faktoren.

Ein scheinbar geringfügiger Zusatz zum Modell des proportionalen Wachstums führt statt zur Lognormalverteilung zu einer asymptotisch nach einem Potenzgesetz abfallenden Häufigkeitsverteilung (Gabaix 1999; Mitzenmacher 2004). Wenn man nämlich die Firmen daran hindert, unter eine minimale Größe zu schrumpfen, entsteht eine Verteilung, bei der die kumulative Verteilungsfunktion F_{cum} für große Firmengrößen s mit $1/s$ abfällt.

Wir realisieren diese Zusatzbedingung, indem wir im oben verwendeten Computer-Programm eine Zeile einfügen, die alle zufällig entstehenden Firmengrößen $s < 1$ auf das Minimum $s_{\min} = 1$ anhebt. Jetzt dauert es etwas länger, bis sich eine stabile Häufigkeitsverteilung herausbildet. Abbildung 4.3 zeigt, dass entsprechend der theoretischen Lösung nach 500 Runden für nahezu alle s gilt: $F_{\text{cum}}(s) \sim 1/s^{\gamma}$ mit $\gamma \approx 1$.

Gabaix (1999) stellte die These auf, dass ein solches Modell das Bevölkerungswachstum in Städten eines Landes erkläre, deren Einwohnerzahlen s oft dem Zipf'schen Potenzgesetz $F_{\text{cum}}(s) \sim 1/s$ folgen (s. Abschnitt 2.2, S. 15).

4.5 Skaleninvarianz

Betrachten wir zuerst noch einmal die in Abschnitt 2.4 (S. 18) erwähnten Größenverteilungen von Jungen eines Alters t_1, die oft angenähert durch eine Normalverteilung mit Mittelwert μ und Standardabweichung σ beschrieben werden können (s. Gleichung 2.8, S. 19). Wir fragen uns nun, wie die Körpergröße der gleichen Jungen zu einem früheren Zeitpunkt $t_0 < t_1$ verteilt gewesen sein könnte. Die einfachste Annahme ist die eines proportionalen Wachstums: Alle Jungen sind in dem Zeitintervall (t_0, t_1) um den Faktor $g > 1$ gewachsen. Der Graph der Verteilung befand sich also für t_0 weiter links auf der x-Achse. Die Dichtefunktion der Größenverteilung zur Zeit t_0 erhält man direkt, wenn man die Skala der x-Achse um den Faktor g vergrößert: $x \to gx$ und die Dichtefunktion $f(gx)$ berechnet. Der Exponent der Dichtefunktion wird dann zu $-(gx-\mu)^2/2\sigma^2 = -(x-\mu/g)^2/2(\sigma/g)^2$. Wir erhalten also eine Normalverteilung mit Mittelwert μ/g und Standardabweichung σ/g. Die Skalenänderung ändert also sowohl die Lage als auch die Form der Gauß'schen Glockenkurve.

Wenn der Logarithmus einer Größe normalverteilt ist (wie es sich z. B. aus dem Gibrat'schen Gesetz ergibt, s. Abschnitt 4.4), dann wird nur der Parameter μ durch eine Skalenänderung $x \to gx$ geändert: $-(\log gx - \mu)^2/2\sigma^2 = -(\log x + \log g - \mu)^2/2\sigma^2$, d. h. $\mu \to \mu - \log g$. Die Glockenkurve von Abbildung 4.2 (S. 43) würde also zu kleineren Werten von $\log x$ hin verschoben werden, ihre durch σ bestimmte Breite aber behalten.

Potenzfunktionen verändern sich bei einer Skalenänderung nur im Normierungsparameter C, der aber wieder auf den alten Wert justiert werden muss, wenn die Potenzfunktion eine Verteilung beschreibt (damit die Summe der relativen Häufigkeiten 1 ergibt). Die Funktion, die die Lotka-Verteilung beschreibt, wird zu $f(gx) = C/(gx)^\alpha = Cg^{-\alpha}/x^\alpha = C'/x^\alpha$. Die Lotka-Verteilung und mit ihr alle Größenverteilungen, die einem Potenzgesetz folgen, sind also invariant gegen Skalenänderung.

Diese *Skaleninvarianz*[6] ist nur Potenzfunktionen eigen, wie man beweisen kann.[7] Sie ist also eine die Potenzfunktionen definierende Eigenschaft. Es liegt daher nahe zu fragen, ob das häufige Auftreten von durch Potenzfunktionen beschriebenen Größenverteilungen etwa auf ihre Skaleninvarianz zurückzuführen ist.

Wir betrachten dazu als Beispiel die von Lotka untersuchte Bibliographie, die er den *Chemical Abstracts* 1907–1916 entnommen hat. Unter der Annahme, dass jeder Autor (wenigstens im Mittel) in den ersten fünf Jahren genau so viel Artikel verfasst hat, wie in den letzten fünf, erhält man wegen der Skaleninvarianz der Lotka-Verteilung für beide Fünf-Jahres-Zeitspannen ein Potenzgesetz mit dem selben Parameter α, wie für alle zehn Jahre zusammen: der Exponent α ist unabhängig von der Größe der Bibliographie. Man kann daher Werte des Lotka-Exponenten für unterschiedlich große Bibliographien direkt vergleichen.

Abraham Bookstein kehrte nun die Argumentation um: Nur statistische Gesetzmäßigkeiten, die robust gegenüber Änderung des betrachteten Zeitintervalls sind, haben überhaupt einen Erkenntniswert (Bookstein 1990). Die Lognormalverteilung der Firmengrößen nach Gibrat ist weitgehend robust, nur der Parameter μ ist zeitabhängig. Am robustesten sind aber Potenzgesetze, die weder Form noch Lage ändern, wenn man größere oder kleinere Zeitspannen untersucht.

Beides gilt jedoch nur unter der Annahme, dass die Produzenten (Autoren, Journale, Firmen), die im erste Zeitraum viel produzieren, auch im zweiten erfolgreich sind. Das erinnert an das Matthäusprinzip – wer hat, dem wird gegeben –, wenn es auch hier nicht einem stochastischen Modell zu Grunde liegt wie beim Yule-Prozess, beim Price'schen Urnenmodell oder beim Gibrat'schen Gesetz.

4.6 Wachstumsmodelle

Wenn jede Frau im Mittel zwei Töchter gebären würde, verdoppelte sich die Menschheit in jeder Generation. Eine konstante Geburtenrate führt zu einem steigenden absoluten Wachstum: es steigt proportional zur Bevölkerungszahl.

Wir suchen nun die mathematische Funktion, die diesem Wachstumsgesetz entspricht. Bezeichne $f(t)$ die Stärke einer Population zum Zeitpunkt t und Δf ihre absolute Zunahme in der Zeiteinheit Δt. Die Zunahme pro Zeiteinheit ist proportional zur Bevölkerungszahl: $\Delta f/\Delta t \sim f$. Die Zeiteinheit Δt kann beliebig klein gewählt werden, so dass der Differenzenquotient $\Delta f/\Delta t$ zum Differentialquotient $\mathrm{d}f(t)/\mathrm{d}t = f'(t)$ wird. Die gesuchte Funktion muss also die Differentialgleichung

$$f'(t) = cf(t)$$

[6] In den letzten Jahren sind viele Arbeiten zu *scale-free networks* erschienen, die gerade dadurch definiert sind, dass ihre *degree distribution* einem Potenzgesetz folgt.

[7] Man zeige als Übung, dass die Exponentialfunktion $\exp(x/x_0)$ nicht skalenivariant ist.

erfüllen. Die einzige Funktion, die proportional zu ihrer Ableitung wächst, ist die Exponentialfunktion:[8]

$$f(t) = f(0)\exp(ct) = f(0)e^{ct} = f(0)e^{t/T}.$$

Die Zeit $T = 1/c$ ist gerade die, in der die Population auf das e-fache anwächst.[9]

Ein annähernd exponentielles Wachstum wird nicht nur für Bevölkerungszahlen beobachtet, sondern auch bei Populationen anderer Lebewesen und – wie deSolla Price (1963 und 1965) zeigte – auch für wissenschaftliche Zeitschriften und die in ihnen publizierten Artikel (s. Abschnitt 2.6, S. 21).

Wissenschaft wächst exponentiell, weil fast jede Lösung eines Problems neue Probleme aufwirft oder die Lösung bereits bekannter Probleme ermöglicht. So kann die Forschung ständig expandieren und tut dies seit Jahrhunderten mit ziemlich konstanter Rate.

Der wachsende Wissensbestand führt zu Spezialisierung, Universalgelehrte gehören schon lange der Vergangenheit an. Auch eine der großen Disziplinen kann ein einzelner Mensch nicht mehr so überblicken, dass er auf allen ihren Teilgebieten forschen könnte. Ständig entstehen neue Spezialgebiete mit neuen Fachzeitschriften.

Wenn wir, wie van Raan (2000), annehmen, dass erstens die Zahl der Fachgebiete exponentiell anwächst und zweitens die Fachgebiete selber ebenfalls exponentiell wachsen (und zwar alle mit gleichem prozentualen Zuwachs pro Jahr), dann lässt sich daraus eine Größenverteilung der Fachgebiete ableiten, die – wie vieles andere in der Wissenschaft – einem Potenzgesetz folgt. Tatsächlich hat van Raan (1990) gefunden, dass die Größenverteilungen von Kozitationsclustern annähernd durch Potenzfunktionen beschrieben werden können.

Es leuchtet ein, dass Größen von Fachgebieten schief verteilt sind: einige große ältere Gebiete stehen vielen kleinen, erst unlängst entstandenen, gegenüber. Das beschriebene Modell von Wachstum und Differenzierung der Wissenschaft erklärt das empirisch gefundene Potenzgesetz vollkommen analog zur von Enrico Fermi (1949) vorgeschlagenen Erklärung der Spektralverteilung der kosmischen Strahlung. Es wurde bereits von dem indischen Physiker und Astronomen S. Naranan (1970) für die Erklärung des Bradford-Gesetzes der Literaturstreuung herangezogen (s. Abschnitt 2.3, S. 16). Einige Zeilen mathematischer Ableitung ergeben eine Potenzgesetz der Größenverteilung von Fachgebieten (Naranan 1971). Es sagt $f(j) = C/j^{1+\alpha/\beta}$ Gebiete der Größe j voraus, wobei α das Wachstum der Zahl der Gebiete charakterisiert ($F(t) \sim \exp \alpha t$) und β das Wachstum der Gebiete selber ($F_i(t) \sim \exp \beta(t - t_i)$, t_i – Entstehungszeit des Gebietes i). Falls beide Wachstumsprozesse gleich schnell ablaufen, gilt also $\alpha = \beta$ und wir erhalten als Exponenten der Potenzfunktion $1 + \alpha/\beta = 2$, wie beim klassischen Fall der Lotka-Verteilung (s. Abschnitt 2.1, S. 13).

Exponentielles Wachstum setzt eine ausreichende Zufuhr von Ressourcen voraus. Beim bisherigen Tempo des Wissenschaftswachstums wird bald eine Schranke erreicht sein, wie schon de Solla Price (1963) in seinem Buch *Little Science, Big Science* selber bemerkte. Er sagte für das von ihm nachgewiesene exponentielle Wachstum der Wissenschaft über drei Jahrhunderte mit Verdopplungszeiten von 15 Jahren ein baldiges Ende voraus, allein schon deswegen, weil wir sonst bald "zwei Wissenschaftler pro Mann, Frau, Kind und Hund der Bevölkerung" hätten (S. 30 der dt. Ausg.). Hinzu kommt, dass besonders in den Natur- und Technikwissenschaften die apparative Ausrüstung immer teurer wird, wodurch das Finanzierungsproblem nicht nur verschärft sondern auch dem Publikum – das dafür Steuern zahlen soll – augenfällig wird. Price diskutiert das Modell des logistischen Wachstums, in dem die begrenzten Ressourcen dadurch berücksichtigt werden, dass das Wachstum nicht nur proportional zur Population $f(t)$ ist, sondern auch zu den noch vorhandenen Ressourcen, die anfangs den Wert G haben und dann mit wachsender Population verbraucht werden: $G - f(t)$. Als Ergebnis erhält man eine Funktion, die anfangs exponentiell steigt, deren Anstieg sich dann verlangsamt und die sich am Ende asymptotisch dem Wert G nähert.[10] Wenn neue Ressourcen erschlossen werden, kann wieder ein logistisches Wachstum bis zum nächsten Grenzwert beginnen.

[8] Mit ihr erfüllen auch die aus ihr abgeleiteten Funktionen mit von e verschiedener Basis die obige Differentialgleichung.

[9] Die Basis $e = 2.718281828459045\ldots$ (Euler'sche Zahl) ist ein unendlicher Dezimalbruch und berechnet sich mit der Definition der Exponentialfunktion $\exp x = \sum_{n=1}^{\infty} x^n/n!$ für $x = 1$.

[10] vgl. auch den Eintrag in Wikipedia: `http://de.wikipedia.org/wiki/Logistische_Funktion`

Kapitel 5

Bibliometrische Indikatoren

5.1 Produktivität

Die Produktivität von Forschenden, ihren Beitrag zum Wissenschaftsfortschritt maß Lotka (1926) anhand der Zahl ihrer Publikationen in einer Fachbibliographie. Er wollte herausfinden, wie häufig Forscher unterschiedlicher Produktivität gewöhnlich auftreten. In zwei ganz unterschiedlichen Bibliographien fand er eine Verteilung der Autoren nach Publikationszahlen, die nach einem Potenzgesetz abfällt (s. Abschnitt 2.1, S. 13).

Heute werden Publikationszahlen von Autoren, Forschungsgruppen und Instituten zu einem ganz anderen Zweck erhoben – es geht nicht um Erkenntnis der Gesetze des Wissenschaftsfortschritts, sondern um die Bewertung der Forschenden, um die Evaluation ihrer Forschung.

Die Konsequenzen von Evaluation können harsch sein – bis hin zur Schließung ganzer Forschungseinrichtungen. Wissenschaft muss sich Evaluation gefallen lassen, wird Forschung doch in vielen Fachgebieten immer teurer. Es ist nicht sinnvoll, wenig produktive Wissenschaftler mit Forschungsmitteln auszustatten, die anderswo besser eingesetzt wären.

Das Anreizsystem von Wissenschaft beruht seit Beginn der Neuzeit auf Reputation, die einzelne Forscher durch publizierte Resultate erwerben (s. Abschnitt 1.1, S. 7). Bibliometriegestützte Evaluation verwendet Indikatoren für Reputation, damit auch Entscheidungsträger, die nicht zur jeweiligen Fachgemeinschaft gehören, Forschungsmittel angemessen vergeben können. *Evaluative Bibliometrie* entwickelt Indikatoren für Produktivität, Wirkung (*impact*) und Kooperation von Forschenden, die zusammen ihre Reputation ausmachen. Sie muss dabei unbeabsichtigte Rückwirkungen auf das Verhalten der Evaluierten möglichst ausschließen.

Ein solch simpler Indikator wie die Zahl von Publikationen in Fachzeitschriften kann Forscher zu Mehrfachpublikation der gleichen Ergebnisse verleiten, auch zu so genannten Salami-Publikationen, in denen nur scheibchenweise der Welt die Resultate kundgetan werden. Man kann sich auch des Eindrucks nicht erwehren, dass manche Koautorschaften als gegenseitige Geschenke vergeben worden sind und keinesfalls auf solchen Beiträgen der Beschenkten zum Artikel beruhen, die eine Autorschaft rechtfertigen (und nicht bereits mit einer Danksagung abgegolten wären, s. Abschnitt 3.7, S. 36). Die ironisch gemeinte Losung *Publish or perish!* (Veröffentliche oder verende!) karikiert die Wirkung des Gebrauchs von Publikationszahlen für die Evaluation der Forschung.

Leichthin vergebene Koautorschaft lohnt sich dann nicht mehr, wenn Publikationen fraktional gezählt werden (vgl. Abschnitt 2.4, S. 18). Dabei muss nicht unbedingt allen k Autoren das gleiche Gewicht $1/k$ zugemessen werden, in einigen Fachgebieten sind gewöhnlich die Erstautoren die, die am meisten zu einem Zeitschriftenaufsatz beigetragen haben. Danach folgt oft der letzte Autor in der Autorenliste, dann die übrigen.[1]

Fraktionale Publikationszahlen haben den Vorteil, additiv zu sein: Zahlen einzelner Autoren können zu denen von Kollektiven addiert werden. Zeitreihen nationaler Publikationszahlen zeigen einen unechten Aufwärtstrend, wenn nicht fraktional gezählt wird, weil generell die internationale Kooperation zunimmt.

Salami- oder Mehrfachpublikationen kann man in gewissem Maße vorbeugen, indem die Publikation in einer Zeitschrift mit deren Reputation gewichtet wird. In hochangesehenen Zeitschriften zu publizieren ist nicht einfach, sie haben eine hohe Ablehnungsrate. Für die evaluative Bibliometrie verschiebt sich damit das

[1] Peter Vinkler (2000, Table 4, S. 608) listet die definierten Anteile einer Publikation auf, die Forscher des Chemischen Forschungszentrums der Ungarischen Akademie der Wissenschaften je nach ihrem Platz in der Autorenliste zugesprochen bekommen. Die Summen für die Forschungsgruppen werden im Institut der Vergabe der Forschungsmittel zugrunde gelegt. Der Algorithmus wurde so zwischen den Gruppenleitern vereinbart.

Problem auf die Messung der Reputation von Journalen. Oft wird sie durch Garfields *Journal Impact-Factor* (JIF) geschätzt, der aber stark von den unterschiedlichen Zitations- und Publikationsgewohnheiten in den einzelnen Fachgebieten abhängt, was Vergleiche über enge Fachgrenzen hinaus verbietet (s. den nächsten Abschnitt 5.2).

5.2 Wirkung (*impact*)

Seitdem es Zitationsdatenbanken gibt, kann die Wirkungsgeschichte einer wissenschaftlichen Publikation ohne großen Aufwand und im Einzelnen nachvollzogen und damit auch quantitativ analysiert werden. Jede Zitierung zeigt Wirkung auf die zitierenden Autoren an, sei es, dass zitierte Resultate von ihnen unmittelbar genutzt werden, oder sei es, dass sie die zitierte Publikation nur bewerten und in den Gang der Forschung einordnen. Jedenfalls erfährt eine Publikation durch die Zitierung Aufmerksamkeit, bekanntermaßen ein rares Gut (s. Abschnitt 2.5, S. 19).

Eine Zitierung kann also in einem Fall viel Wirkung bedeuten, im anderen ganz wenig. Neuerdings wird versucht, Zitierungen anhand der Zitierungskontexte automatisch zu klassifizieren.[2]

Stark beachtete Publikationen müssen irgendeine Qualität aufweisen, die Aufmerksamkeit erregt. Der Umkehrschluss von Qualität auf Beachtung gilt nicht zwingend, so dass wenig zitierte Arbeiten nicht von minderer Qualität sein müssen.[3] Auch originelle, methodisch anspruchsvolle und gut formulierte Publikationen müssen nicht notwendig viel Beachtung finden. Die hängt stark davon ab, wie viele Forscher sich gerade für die jeweilige Thematik interessieren.

Bei der Konstruktion von Wirkungsindikatoren von Publikationen und bei ihrer Anwendung muss daher berücksichtigt werden, dass Forschungsrichtungen ganz unterschiedlich intensiv betrieben werden, in ihnen ganz unterschiedlich viel publiziert wird und damit auch ganz unterschiedlich viel zitiert werden kann. Die mittlere Länge der Referenzenlisten zitierter Quellen bestimmt, wieviel Zitierungen eine Publikation im Forschungsgebiet im Mittel erwarten kann.

Ein einfaches Beispiel soll dies verdeutlichen. In einem Fachgebiet erscheinen pro Jahr $f = 500$ Artikel. Wir betrachten nun die Chance eines Artikels aus einem Doppeljahrgang von $2f = 1000$ Artikeln im darauffolgenden dritten Jahr in einem der wiederum $f = 500$ Artikel des Fachgebiets zitiert zu werden.[4] Wir nehmen an, im Mittel würden im dritten Jahr in jedem der $f = 500$ Artikel $q = 6$ Quellen aus dem vorangegangenen Doppeljahrgang des Fachgebiets zitiert. Dann bekommt jede dieser $2f = 1000$ Quellen im Mittel $fq/2f = q/2 = 3$ Zitierungen. Wir erhalten also im Beispiel eine Zitierrate $q/2$, die unabhängig von der jährlichen Artikelzahl f ist und nur von der mittleren Zahl zitierter Quellen q abhängt. In großen Fachgebieten werden jedoch oft mehr Quellen pro Artikel zitiert als in kleinen, weil man mehr zitieren kann und muss. Deswegen ist z. B. die mittlere Zitationszahl von biomedizinischen Artikeln höher als von mathematischen oder physikalischen.

Ist eine Publikation doppelt so oft zitiert worden, wie eine andere, hat sie auch das doppelte Maß an Aufmerksamkeit erzielt. Sie ist deswegen aber nicht unbedingt doppelt so wichtig für den Fortschritt im jeweiligen Fachgebiet. Es gilt hier wieder, dass sinnvoll nur Gleiches mit Gleichem verglichen werden kann. Die Beachtung von im selben Heft einer Zeitschrift erschienenen Artikel kann jedoch mit einiger Berechtigung unmittelbar an den jeweils bis zu einem Zeitpunkt erzielten Zitierungen abgelesen werden.

Aber auch in solchen Fällen können gleiche Zitierungszahlen unterschiedlicher Wirkung entsprechen, je nach dem an wie vielen Zitierungen der Autor (bzw. die Autoren) der zitierten Publikation beteiligt ist (bzw. sind), wie viele Zitierungen von nahen Kollegen kommen und wie viel internationale Beachtung die Publikation gefunden hat. Selbstzitierungen bei der Konstruktion von Wirkungsindikatoren einfach wegzulassen, ist aber methodisch fragwürdig wegen der in vielen Fachgebieten immer noch steigenden Zahl von Autoren pro Publikation. Wenn in einem Aufsatz mit fünf Autoren ein anderer zitiert wird, bei dem einer der fünf Koautor von vier anderen Forschern ist, wäre es nur zu 1/25 berechtigt, dies als Selbstzitierung anzusehen.[5]

Der Ausweg der Bibliometrie aus einschränkenden Erwägungen dieser Art ist ja immer der Rückzug auf die Statistik. Ist auch in jedem einzelnen Fall die Sache differenziert zu betrachten, für genügend große Bibliographien ist die Verwendung zitationsgestützter Indikatoren für die Wirkung doch sinnvoll, weil sich individuelle Unterschiede herausmitteln. Vergleicht man aber einzelne Publikationen, dann

[2] s. z. B. `http://www.eerqi.eu`

[3] ein besonders von Wolfgang Glänzel immer wieder betontes Argument

[4] Das sind gerade die beim *Journal Impact-Factor* (s. u.) verwendeten Publikations- und Zitationsfenster.

[5] vgl. Schubert, Glänzel und Thijs (2006)

müssen alle oben angeführten Einschränkungen berücksichtigt werden.

Journal Impact-Factor

Mit dem Ziel, ein Maß für die Reputation von Zeitschriften zu konstruieren, haben Garfield und Sher (1963) den *Journal Impact-Factor* (JIF) erfunden, welcher als mittlere Zitierungszahl von Artikeln zweier Jahrgänge einer Zeitschrift im darauffolgenden Jahr definiert ist. Er wird jährlich für alle im *Web of Science* indexierten Journale berechnet und ist in den *Journal Citation Reports* einsehbar.[6]

Ein Mangel des *Journal Impact-Factors* ist neben der oben erwähnten Abhängigkeit von den Zitations- und Publikationsgewohnheiten des Fachgebietes, dass er als arithmetisches Mittel einer schiefen Verteilung keine Vorhersagekraft für die Zitierung eines einzelnen Artikels hat: die meisten Artikel eines Doppeljahrgangs eines Journals werden im Folgejahr wenig oder gar nicht zitiert, einige hochzitierte beeinflussen den Wert des *Journal Impact-Factors* stark, haben jedoch eine weit höhere Zitierrate, als der JIF angibt. In seltenen Fällen vervielfachen einzelne extrem hochzitierte Aufsätze den JIF sogar (Czerwon und Havemann 1991). Meistens ist sein Wert über die Jahre jedoch ziemlich stabil. Dass Schwankungen gering bleiben, liegt auch daran, dass sich die Publikationsfenster zweier aufeinanderfolgender *Journal Impact-Factors* eines Journals überlappen. Sieht man als grobes Modell die Zitationszahlen einzelner Jahrgänge als zufällige Stichproben aus ein und der selben Grundgesamtheit an, ergibt sich nach dem zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung auch für schief verteilte Grundgesamtheiten, dass die *Journal Impact-Factors* als Probenmittelwerte normalverteilt sind. Das vor allem erklärt die Stabilität des *Journal Impact-Factors*: Ausreißer sind selten (Glänzel 2008).

Es gab unzählige Vorschläge für verbesserte Journal-*Impact*-Indikatoren, die sich aber sämtlich nicht gegen Garfields *Journal Impact-Factor* durchsetzen konnten. Am häufigsten werden noch Varianten des *Impact-Factors* mit verlängerten Publikations- und Zitationsfenstern für Analysen benutzt. Ein größeres Publikationsfenster reicht zeitlich weiter zurück, was den Verhältnissen in Fachgebieten mit langsamer Kommunikation angemessener ist als die zwei berücksichtigten Vorjahre bei Garfields JIF. Nicht in allen Gebieten wird das Maximum der Zitationszahlen eines Jahrgangs in den beiden darauffolgenden Jahren erreicht (s. Abschnitt 2.6, S. 21).

Auch der Prestige-Indikator für Journale, den Pinski und Narin (1976) vorgeschlagen haben, war als Alternative zum Garfield'schen *Journal Impact-Factor* gedacht, der auch die unterschiedlichen Zitationsgewohnheiten in den Fachgebieten ausgleicht (s. Abschnitt 3.2, S. 27). In Anschluss an Googles Erfolg mit dem *PageRank*-Algorithmus wird das beiden Indikatoren zugrundeliegende Prinzip aus der sozialen Netzwerkanalyse, die Zitierungen (bzw. die Links) mit dem Prestige der zitierenden Zeitschrift (bzw. der verlinkenden Webpage) selber zu wichten und dann mit einem Rückkopplungsalgorithmus zu iterieren (*bootstrapping*) neuerdings auch breiter für die Wichtung von Journalen angewandt.[7]

Relative Citation Rates (RCR)

Der Einfluss der in den Fachgebieten unterschiedlichen Zitationsgewohnheiten auf von Artikeln eines Autors oder einer Gruppe von Autoren erzielte Zitierraten kann mittels des *Journal Impact-Factors* (bzw. eines *Impact*-Faktors mit verändertem Publikations- und Zitationsfenster) auf folgende Weise herausgerechnet werden (Schubert und Braun 1986).[8] Man berechnet zum einen die mittlere Zahl der im Zitationsfenster beobachteten Zitationen der im Publikationsfenster erschienen Artikel, die *MOCR* (*mean observed citation rate*) genannt wird. Als zweiter Indikator wird eine mittlere erwartete Zitierrate *MECR* (*mean expected citation rate*) bestimmt. Dabei geht man von den Publikationskanälen der Autoren aus, d. h. von den Journalen, in denen sie publizieren. Als erwartete Zitierrate eines Journals wird gerade sein *Impact*-Faktor verwendet, berechnet mit den selben Zeitfenstern wie *MOCR*. *MECR* ist dann das gewichtete Mittel der *Impact*-Faktoren der Publikationskanäle. Autoren unterschiedlicher Fachgebiete können so mittels der *Relative Citation Rate* $RCR = MOCR/MECR$ verglichen werden.

Für den Vergleich von Autoren (oder Autorengruppen) innerhalb des gleichen Fachgebietes zeigt der $MECR$-Indikator, wie anspruchsvoll sie jeweils bei der Wahl ihrer Publikationskanäle

[6] Zu beachten ist, dass im Zähler des JIF zwar Zitationen aller Artikel im jeweils aktuellen Jahr (Zitationsfenster) berücksichtigt werden, im Nenner jedoch lediglich die Zahl der so genannten *citable items* in den beiden Vorjahren (Publikationsfenster) erscheint (*Citable items* sind Dokumente der Typen *article*, *letter*, *note* und *review*).

[7] `http://www.scimagojr.com`

[8] s. a. das Vorlesungsskript von Wolfgang Glänzel (2003, S. 66)

sind und am $MOCR$-Indikator lässt sich ablesen, ob sie die in ihren Journalen übliche Zitationsrate im Mittel übetreffen oder nicht. In einem $MECR$-$MOCR$-Diagramm findet man die betrachteten Autoren(-Gruppen) als Punkte entlang der Geraden $MOCR = MECR$ (die also $RCR = 1$ entspricht). Autoren oberhalb der Geraden sind im allgemeinen besser als der Durchschnitt in ihren Journalen, die darunter schlechter.

Hirsch-Index

Ein viel diskutierter zitationsbasierter Indikator wurde von dem in Kalifornien tätigen Physiker Jorge E. Hirsch (2005) vorgeschlagen. Der vom Autor als h-Index bezeichnete Indikator ist für eine Bibliographie (eines Autors, eines Instituts, eines Journals...) auf höchst einfache Weise definiert: sie hat den Index h, wenn h ihrer Publikationen mindestens h-mal zitiert worden sind und alle anderen weniger oft.

Hirsch hatte vor allem Lebensbibliographien von Autoren mit allen jemals erhaltenen Zitierungen im Blick. Im *Web of Science* kann man sich den h-Index beliebiger Bibliographien von Zeitschriftenaufsätzen anzeigen lassen.[9] Der neue Indikator wurde auch deshalb so schnell populär, weil er so einfach zu bestimmen ist: es genügt, die Bibliographie absteigend nach der Zitationszahl zu sortieren und die letzte Publikation in der Rangliste zu suchen, deren Rangzahl noch nicht die Zitationszahl übersteigt.

Wird die höchstzitierte Publikation eines Autors immer wieder zitiert, erhöht das den h-Index überhaupt nicht. Der Matthäus-Effekt wird also wenigstens gedämpft: nur eine breite Spitze vielzitierter Publikationen wird mit hohem Hirsch-Index belohnt. Die Unabhängigkeit von den hohen Zitationszahlen kann man sich am Diagramm in Abbildung 5.1 verdeutlichen. Falls der Spitzenreiter (mit 62 Zitierungen) noch zehn Zitierungen dazubekäme, bliebe der Hirsch-Index dennoch $h = 18$.

Hirsch selber stellte fest (ebenda, S. 16572), dass die prominentesten Autoren in den Biowissenschaften höhere h-Werte erreichen (bis $h = 191$) als Autoren in der Physik (bis $h = 110$). Dies zeigt die Abhängigkeit auch des Hirsch-Indexes von den Zitationsgewohnheiten in Disziplinen und Fachgebieten: längere Referenzenlisten biowissenschaftlicher Artikel führen zu höheren Zitationszahlen als in der Physik.

[9] Er wurde auch in die Zitationsanzeige der SPIRES-Datenbank der Hochenergiephysik einbezogen, und zwar noch bevor der als *arXiv*-Preprint publik gemachte Artikel von Hirsch als Zeitschriftenaufsatz erschien war (vgl. `http://www.slac.stanford.edu/spires/hep`).

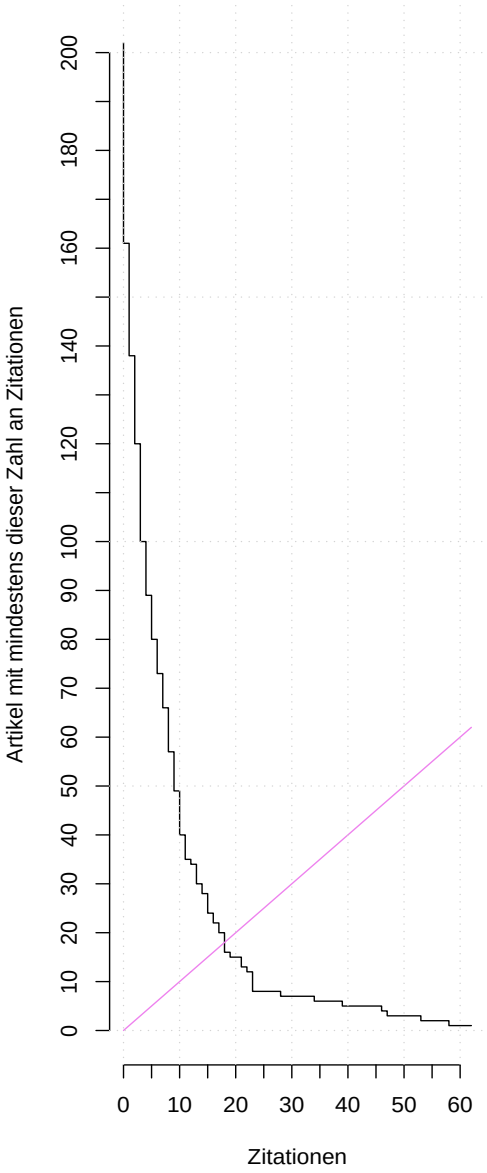


Abbildung 5.1: Verteilung der 2001 und 2002 in *Scientometrics* publizierten Artikel nach ihren Zitationszahlen bis zum 9. Juli 2008 (Datenquelle: *Web Of Science*). Die Gerade $y = x$ schneidet die Treppenkurve der kumulierten Artikelzahlen beim Hirsch-Index $h = x = y = 18$.

Auch für den Hirsch-Index wurden bereits unzählige Varianten entwickelt, die vielleicht für bestimmte Zwecke jeweils besser geeignet sind als das Original, das ihnen allen aber eins voraus hat – seine Einfachheit.

5.3 Kooperation

Zusammenarbeit in der Forschung äußert sich als Koautorschaft und wird dadurch bibliometrisch erfassbar. Forschungskooperation nimmt zu. Diese Tendenz zu mehr Kooperation – auch über Instituts-, Länder- und Fachgebietsgrenzen hinweg – zeigt sich in den Metadaten von Zeitschriftenaufsätzen: Autoren- und Adressenliste der Artikel werden länger, Aufsätze mit nur einem Autor sind bereits in vielen Fachgebieten in der Minderheit, in manchen sogar selten. Die Verteilungen von Artikeln nach der Zahl der Autoren verschieben sich zu höheren Autorzahlen, wie die Abbildungen 5.2–5.4 am Beispiel von fünf bibliometrisch orientierten Journalen verdeutlichen (welche schon in Abschnitt 3.7, S. 36, als Beispiel dienten). Die Tendenz zu mehr Kooperation ist – über das gesamte vergangene Jahrhundert – so eindeutig, dass sie erklärt werden muss.[10]

Welche der Ursachen für Kooperation haben sich verstärkt? Welche der Hindernisse sind geringer geworden? Bei den verbesserten Bedingungen für Zusammenarbeit herrscht weitgehend Übereinstimmung. Insbesondere internationale Kooperation wird immer leichter: Englisch setzt sich als Wissenschaftssprache immer mehr durch, der Eiserne Vorhang ist verschwunden, Flüge haben sich verbilligt, und das Internet macht die Kommunikation einfach, billig und schnell. Gerade das Internet wurde jedoch von den Wissenschaftlern selber geschaffen und zu ihrem mittlerweile wichtigsten Kommunikationsinstrument gemacht (bevor es von der Wirtschaft entdeckt wurde). Es kann nicht nur an dem Wegfallen von Barrieren oder der besonderen Förderung liegen, wenn Kooperation auf allen Ebenen immer mehr zugenommen hat und noch weiter zunimmt. Es muss tieferliegende Ursachen für diese säkulare und globale Tendenz geben. Bei der Diskussion der Ursachen sind die Meinungen vielfältiger.

Als kognitive Ursache für die Zunahme von Zusammenarbeit sehe ich an, dass die Ertragsrate der Forschung sinkt (Rescher 1982). Aus der ständig wachsenden Masse an vorhandenem Wissen können und müssen im Mittel für das jeweilige Forschungsproblem immer mehr Fakten, Theorien und Methoden zur Verfügung stehen, was einzelne Forscher und Institutionen nicht mehr liefern können (Hicks und Katz 1996, S. 41).

[10] Die folgenden Absätze zur Forschungskooperation hat der Autor bereits 2002 in einer für das Bundesministerium für Forschung und Technologie verfassten Expertise verwendet (vgl. Online-Publikation auf `http://www.sciencepolicystudies.de`).

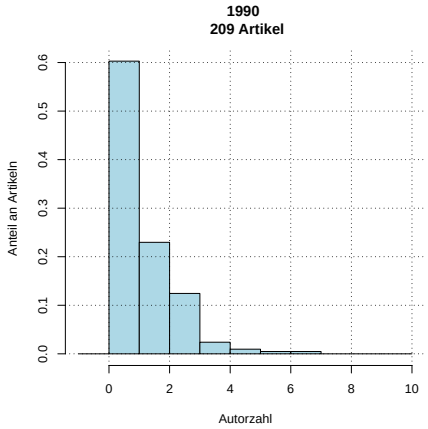


Abbildung 5.2: Histogramm 1990 der *articles* in fünf informationswissenschaftlichen Journalen nach der Autorzahl (s. Text, Datenquelle: *Web of Science*)

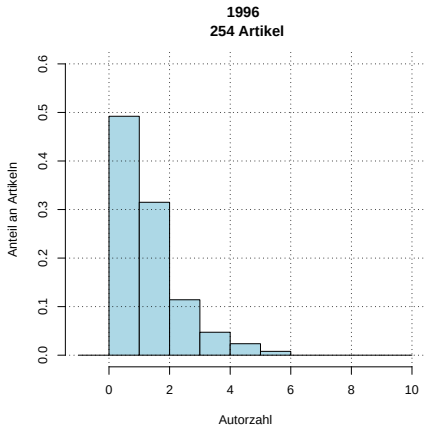


Abbildung 5.3: Dasselbe 1996

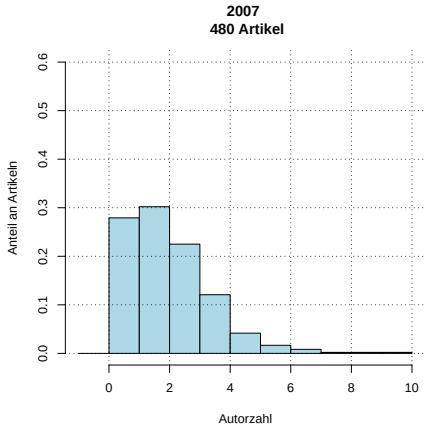


Abbildung 5.4: Dasselbe 2007 (zusätzlich zu den fünf Journalen sind hier auch 31 Artikel aus dem *J. Informetr.* berücksichtigt)

Beim Sozialen würde ich zuallererst darauf verweisen, dass sich in gesellschaftlich (inklusive wirtschaftlich) relevanten – und daher gut dotierten Forschungsrichtungen – ausgehend von den internationalen Spitzeneinrichtungen ständig der Standard und damit der Konkurrenzdruck erhöht, was zur Konzentration der Mittel führt, und dies nicht nur auf der Ebene der Forscher, sondern auch auf nationaler Ebene.

Wie nun Kooperation messen, einen Indikator der Koautorschaft konstruieren? Wir haben es hier wieder mit einer schiefen Verteilung zu tun (wenn auch ihre Schiefe abnimmt): vielen Artikeln mit wenigen Autoren stehen einige wenige Artikel mit relativ vielen Autoren gegenüber. Diese Ausreißer machen das arithmetische Mittel, die mittlere Autorenzahl – besonders bei kleinen Stichproben – genauso instabil, wie die hochzitierten Aufsätze den *Journal Impact-Factor* (vgl. Abschnitt 5.2, S. 48). Bei Fachbibliographien mit einigen hundert Artikeln pro Jahr wird jedoch auch durch die mittlere Autorenzahl die Tendenz zu mehr Kooperation deutlich sichtbar, wie man an Abbildung 5.5 sieht.

Die Abbildung 5.5 zeigt als zweite Kurve die Zeitreihe des Kooperationskoeffizienten C_c (*Collaborative Coefficient*), den Ajiferuke, Burrell und Tague (1988) vorgeschlagen haben. Sie gingen dabei davon aus, dass sich in Fachgebieten mit mehr Kooperation die Autoren im Mittel kleinere Anteile von Artikeln anrechnen können, wenn ihre Produktivität fraktional mit $1/k$ bei k Autoren eines Artikels gemessen wird (vgl. Abschnitt 5.1, S. 47).

Der mittlere Kehrwert der Autorzahlen in einer Bibliographie von n Artikeln ist

$$\left\langle \frac{1}{k} \right\rangle = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_i}.$$

Als Mittelwert von Zahlen, die ≤ 1 sind, kann auch $\left\langle \frac{1}{k} \right\rangle$ den Wert 1 nicht überschreiten. Ausreißer ($k_i \gg 1$) beeinflussen seine Größe nur wenig, auch in kleinen Stichproben. Um einen Indikator für Kooperation zu bekommen, subtrahieren Ajiferuke *et al.* $\left\langle \frac{1}{k} \right\rangle$ von 1:

$$C_\mathrm{c} = 1 - \left\langle \frac{1}{k} \right\rangle = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k_i}. \qquad (5.1)$$

$C_\mathrm{c} = 0$ ergibt sich nur, wenn alle $k_i = 1$, d. h., wenn keine Kooperation vorliegt. $C_\mathrm{c} = 1$ kann nie erreicht werden.

Der Kooperationskoeffizient eines einzelnen Autors quantifiziert seine Neigung zur Zusammenarbeit in der Forschung. Er hängt unmittelbar mit seiner fraktional gezählten Publikationsproduktivität $f = \sum_{i=1}^{n} 1/k_i$ und mit seiner normal gezählten Produktivität n zusammen: $C_\mathrm{c} = 1 - f/n$.

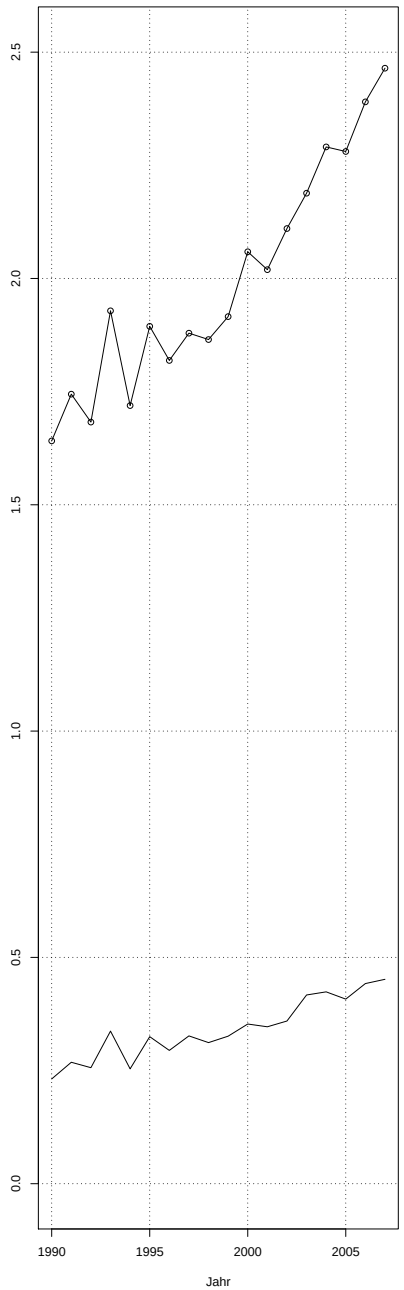


Abbildung 5.5: 1990–2007: Mittlere Autorenzahl (Punkte mit Linien oben) und Kooperationskoeffizient (Kurve unten) für fünf (2007 sechs) informationswissenschaftliche Journale (s. Text)

Gegen die Verwendung von f als Indikator der Produktivität einzelner Autoren wird zuweilen eingewandt, dies benachteilige kooperative Au-

toren. Wenn neben f für die Produktivität auch C_c für die Kooperativität eines Autors angegeben wird, erhält man ein vollständigeres Bild, das allerdings nicht zu einer Rangliste kondensiert werden kann.

5.4 Zitationsverhalten

Aufsätze in wissenschaftlichen Zeitschriften haben heute nicht nur mehr Autoren als vor zehn oder zwanzig Jahren, in ihnen werden auch mehr Quellen zitiert als früher. Die Abbildungen 5.6 bis 5.9 machen deutlich, wie sich die Verteilung der Zahlen zitierter Quellen in den fünf schon im vorigen Abschnitt betrachteten bibliometrisch ausgerichteten Journalen von 1990 bis 2007 zu höheren Referenzenzahlen pro Aufsatz verschoben hat. Abbildung 5.6 zeigt Zeitreihen von Median und von arithmetischem wie geometrischem Mittel der Zahl der Aufsätze.

Das geometrische Mittel als Maß der zentralen Tendenz ist als die n-te Wurzel aus dem Produkt der betrachteten n Zahlen x_i definiert. Be-

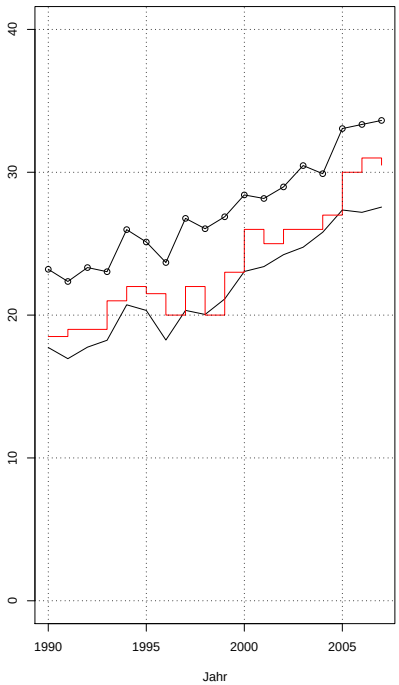


Abbildung 5.6: 1990–2007: Arithmetisches Mittel der Referenzenzahl (Punkte mit Linien oben), entsprechendes geometrisches Mittel (Kurve unten) und Median (rote Treppenkurve) für fünf (2007 sechs) informationswissenschaftliche Journale (s. Text)

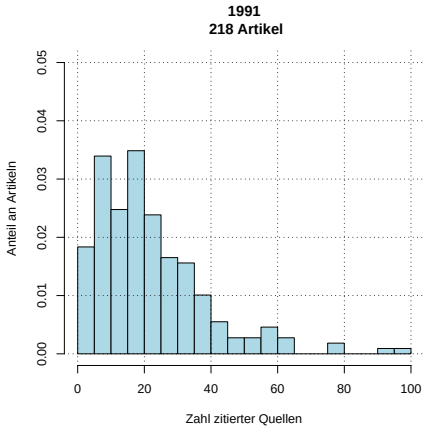


Abbildung 5.7: Histogramm 1991 der *articles* in fünf informationswissenschaftlichen Journalen nach der Referenzenzahl (> 0, auf der y-Achse werden die mittleren Anteile in den Klassen 1–5, 6–10, ... Referenzen angezeigt, s. Text, Datenquelle: *Web Of Science*)

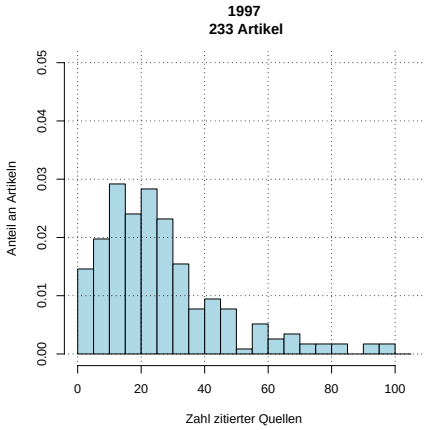


Abbildung 5.8: Histogramm 1997 (wie oben)

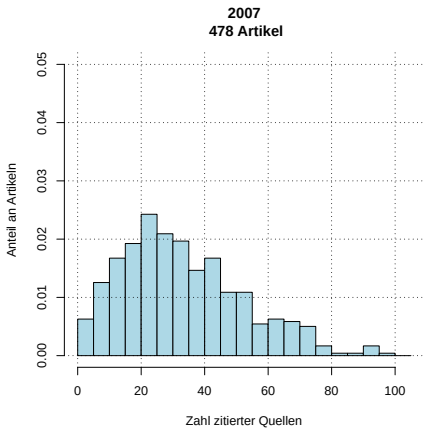


Abbildung 5.9: Histogramm 2007 (wie oben)

Tabelle 5.1: *Journal Impact-Factors* der fünf informationswissenschaftlichen Zeitschriften 1995, 1999 und 2007. Datenquelle: *Journal Citation Reports* des *Web of Science*, Dezember 2008

Journal	1995	1999	2007
Inform. Proc. & Man.	0.58	0.73	1.50
Scientometrics	0.44	0.93	1.47
JASIS[T]	1.16	1.33	1.44
J. Doc.	0.93	1.60	1.31
J. Inform. Science	0.47	0.66	1.08

rechnet wird zumeist erst der Logarithmus dieser Größe:

$$\log(\prod_{i=1}^{n} x_i)^{1/n} = \frac{1}{n}\sum_{i=1}^{n} \log x_i,$$

also das arithmetische Mittel der Logarithmen, das als Argument der Exponentialfunktion dann das geometrische Mittel ergibt. Hieran wird deutlich, warum das geometrische Mittel für schiefe Verteilungen – wie die der Referenzenzahlen – stabiler ist als das arithmetische: die Ausreißer werden durch den Logarithmus "zurückgeholt" (s. Abbildung 5.6).

Alle drei Kurven zeigen die gleiche deutliche Tendenz hin zu mehr zitierten Quellen pro Aufsatz. Mit diesem veränderten Zitierverhalten steigen auch Zitierraten, wie der *Journal Impact-Factor* der fünf Journale (Tabelle 5.1). Die Ursachen dieser Tendenz sind noch nicht geklärt.

Persson, Glänzel und Danell (2004) fanden heraus, dass beide Tendenzen, die zu mehr Referenzen und die zu mehr Autoren, zusammenhängen: in Artikeln mit mehr Autoren werden im Mittel auch mehr Quellen zitiert, ein einleuchtendes Resultat. Aber allein auf die im Vergleich zu Einzelautoren umfangreichere Literaturkenntnis von Autorenkollektiven kann die größere Quellenzahl sicherlich nicht zurückgeführt werden. Daher werden die Änderungen im Zitierverhalten neuerdings detaillierter untersucht, nicht nur anhand der Referenzenzahlen, sondern auch anhand der Altersstruktur der zitierten Quellen und ihrer Verteilung auf fachlich näher oder ferner liegende Zeitschriften.

Die Verteilung der Referenzen nach ihrem Alter hängt stark vom Fachgebiet ab. In sich schnell entwickelnden Gebieten mit schneller Kommunikation kann man einen höheren Anteil jüngerer Quellen erwarten. Als Indikator für das zeitliche Zitierverhalten der Autoren einer Zeitschrift wird in den *Journal Citation Reports* des *Web of Science* die zeitliche Verteilung der Referenzen durch die Größe *Citing Halflife* charakterisiert. Sie ist als der auf Zehnteljahre genaue Median des Alters aller Referenzen definiert. Da jedoch in den Referenzen der Zeitpunkt der Publikation der zitierten Quelle nur als Jahreszahl angegeben ist, ermittelt man die Stelle nach dem Komma durch Interpolation (unter Annahme der zeitlichen Gleichverteilung der zitierten Quellen jedes Jahres und der Zitierung am 1. Januar des Auswertungsjahres).[11]

Hat sich nun die *Citing Halflife* mit den längeren Referenzenlisten ebenfalls verlängert? Anders gefragt: sind die zusätzlich zitierten Quellen gerade die älteren? Die *Citing Halflifes* der fünf oben betrachteten Journale haben sich von 1999 bis 2007 nicht wesentlich verändert. Zu prüfen wäre, ob lange Referenzenliste im Vergleich mit kurzen eher auch ältere Quellen enthalten.

[11] Die Bezeichnung als Halbwertszeit rührt von der früh bemühten Analogie zwischen Literaturalterung und radioaktivem Zerfall her, die hier auf retrospektive Analysen der zeitlichen Verteilung der Referenzen übertragen wird (s. Abschnitt 2.6, S. 21).

Kapitel 6

Anwendungen

6.1 Evaluative Bibliometrie

Die einfachen und naheliegenden Dinge sind schon erforscht. Forschung wagt sich an immer komplexere Gegenstände, die unserer Alltagswelt immer ferner liegen, im Großen wie im Kleinen. Dafür benötigt man große Computer, Forschungs- und Raumschiffe, Tele- und Mikroskope, nicht zu reden von den riesigen Teilchenbeschleunigern der Hochenergiephysik.[1]

Wissenschaft wird aber nicht nur immer teurer, sie wächst seit Jahrhunderten schneller als die Wirtschaftskraft der Länder und verbraucht daher immer größere Anteile des Bruttosozialprodukts (s. Abschnitt 2.6, S. 21).

Die steigenden Aufwendungen müssen einmal ganz allgemein gegenüber der Gesellschaft gerechtfertigt werden. Zum anderen muss jede einzelne Entscheidung bei der Vergabe von Projektmitteln, bei der Ausstattung von Forschungsinstituten sehr sorgfältig getroffen werden, damit die Finanzmittel möglichst wirkungsvoll – mit dem größten Nutzen für die Gesellschaft – eingesetzt werden. Das verschlingt Zeit von Experten, die die Projekte und Institute evaluieren.

Hier nun soll Bibliometrie helfend einspringen. Die Entscheidungsträger in den Ministerien und Forschungsinstitutionen hoffen, durch sie weniger aufwendig und weniger subjektiv Forschungsleistungen bewerten zu können, damit vor allem die Forscher gut ausgestattet werden, welche schon ihre Fähigkeiten bewiesen haben. Bibliometrische Indikatoren für Produktivität, Wirkung und Kooperation von Wissenschaftlern sind gerade auch für evaluative Zwecke entwickelt worden (s. voriges Kapitel).

Wenn Gelder und Stellen nach Publikations- und Zitationszahlen verteilt werden, so passen Forscher ihr Publikationsverhalten entsprechend an: sie publizieren mehr und versuchen, ihre Aufsätze in angesehenen Journalen unterzubringen. Mehr von erhaltenen Forschungsresultaten zu veröffentlichen, ist eigentlich nicht schlecht – wenn höhere Publikationszahlen nicht durch mehrfaches Publizieren ein und desselben Resultats oder durch ungerechtfertigte Koautorschaft an Artikeln erreicht werden.

Evaluative Bibliometrie muss Indikatoren entwickeln, die erwünschte Anpassungen der Evaluierten bewirken und unerwünschte möglichst verhindern. Ungerechtfertigte Koautorschaften können durch fraktionales Zählen der Publikationen, das sich aber noch nicht durchgesetzt hat, zurückgedrängt werden (vgl. Abschnitt 5.1, S. 47).

Der Taktik, dasselbe mehrfach zu publizieren, kann man dadurch entgegenwirken, dass Publikationen mit einem Maß für das Ansehen der Zeitschrift gewichtet werden. Es gibt Institute, die bei der Evaluation ihrer Mitarbeiter schon längere Zeit so verfahren und dafür den *Journal Impact-Factor* (JIF) benutzen. Es leuchtet ein, dass ein Artikel in der angesehen Zeitschrift *Nature* (2007 JIF = 28.751) für die Autoren einen größeren Erfolg darstellt als z. B. ein Aufsatz in *Scientometrics* (2007 JIF = 1.472). Dennoch würde ich den *Nature*-Artikel nicht zwanzigfach höher einschätzen. Der JIF von *Nature* als multidisziplinärer Zeitschrift wird durch die Zitationen naturwissenschaftlicher Artikel bestimmt, die im Mittel weit öfter zitiert werden als sozialwissenschaftliche. Die Abhängigkeit des *Journal Impact-Factors* von den Zitationsgewohnheiten des Fachgebiets macht ihn ungeeignet für die Messung der Reputation von Journalen (vgl. S. 49).

Zitationsbasierte Indikatoren für Forschungsleistung haben gegenüber den rein publikationsbasierten einen Nachteil: es dauert, bis Publikationen zitiert werden. Man misst mit ihnen also Leistungen, die schon einige Jahre zurückliegen. Bei der Evaluation einzelner ist ihr gesamtes Le-

[1] Die Verteuerung der Wissenschaft kann man daran ablesen, dass die Aufwendungen insbesondere für Forschungsgeräte schneller steigen als die Zahl der Publikationen. Das haben schon de Solla Price (1963, S. 104 der dt. Ausg. 1974) sowie Nalimov und Mul'čenko (1969, S. 41) nachgewiesen.

benswerk von Belang, bei Gruppen in einem Institut eher die aktuelle Leistung.

Bibliometriegestützte Evaluation wird dann am ehesten akzeptiert, wenn ihre Kriterien und Ergebnisse mit den Betroffenen diskutiert werden. Sie breitet sich immer mehr aus und hat auch schon die Sozialwissenschaften erreicht.

6.2 *Information Retrieval*

Praktische Anwendungen bibliometrischer Methoden in Bibliotheken waren und sind selten. Am ehesten schien sich noch eine Bradford-Analyse als Hilfsmittel für die Auswahl von Zeitschriften-Abonnements für Spezialinstitute zu eignen (s. Abschnitt 2.3, S. 16). Eine solche Erwerbungspraxis hat sich jedoch nicht eingebürgert und wird in Zeiten von Bibliotheks-Konsortien und *Open Access* zunehmend obsolet.

Die von Bradford gefundene schiefe Verteilung von Zeitschriften nach der Zahl von Aufsätzen zu einer Thematik findet aber beim *Information Retrieval* (IR) Anwendung. Es erweist sich als sinnvoll, bei der Anzeige der Ergebnisse einer textbasierten Recherche in einer Datenbank wissenschaftlicher Zeitschriftenliteratur die Treffer in den Kernzeitschriften, d. h. den mit den meisten relevanten Aufsätzen, zuerst anzuzeigen. Diese *Bradfordising* genannte Methode erhöht die *precision* der in der Rangliste oben stehenden Aufsätze (White 1981; Mayr 2008).

Google Scholar zeigt zu jeder Trefferliste von wissenschaftlichen Web-Dokumenten die für das Thema wichtigsten, d. h. produktivsten Autoren an. Klickt man auf den Namen eines dieser Autoren, kann man ebenfalls eine höhere *precision* der Treffer erzielen, weil die produktivsten Autoren eines Gebietes nach Lotka sich genauso deutlich von den weniger produktiven absetzen, wie Bradfords Kernzeitschriften von den übrigen Zeitschriften (s. Abschnitt 2.1, S. 13).

Bibliometrische Konzepte sind auch Grundlage von anderen Diensten für Nutzer bibliographischer Datenbanken. Das älteste Beispiel dafür ist wohl die Anzeige von *Related Records* in der CD-ROM-Version des *Science Citation Index* (SCI), wobei *related* hier heißt, dass die angezeigten Aufsätze am stärksten mit dem Ausgangsdokument bibliographisch gekoppelt sind (s. Abschnitt 3.4, S. 31).

Googles PageRank-Algorithmus kann als Weiterentwicklung der Geller'schen Variante des Pinski-Narin-Algorithmus angesehen werden (s. Abschnitte 3.2 und 3.3, S. 27 und S. 30). Obwohl sich die *Google*-Gründer Brin und Page (1998) auf die Analogie von Weblinks und Zitierungen beziehen, zitieren sie aber weder Geller (1978) noch Pinski und Narin (1976). Diese werden aber sehr wohl in der Literatur zum *Information Retrieval* im Web rezipiert. Kurz nach Brin und Page schlug Jon M. Kleinberg (1999) eine etwas andere Methode für die Gewichtung der Ergebnisse von Web-Recherchen vor, die aber ähnliche Prämissen benutzt wie die von ihm auch zitierten bibliometrischen Vorläufer.

6.3 Wissenschaftsforschung

Bibliotheks- und Informationswissenschaft als Forschunggebiet, das auf die optimale Informationsversorgung von Wissenschaft, Wirtschaft und Öffentlichkeit zielt, kann nicht ohne quantitative Analyse der Informationsströme auskommen. d. h. nicht ohne Bibliometrie und Informetrie. Bibliometrisches und informetrisches Wissen findet aber nicht nur Anwendung bei der Lösung bibliotheks- und informationswissenschaftlicher Probleme, es bildet einen wesentlichen Teil der empirischen Basis dieses Forschunggebietes, bibliometrische Modelle und Regelhaftigkeiten sind Teil der bibliotheks- und informationswissenschaftlichen Theorie. Bibliometrie und Informetrie sind (sich überlappende) Teilgebiete der Bibliotheks- und Informationswissenschaft (s. Abschnitt 1.3, S. 9).

Bibliometrische Methoden können aber auch bei der Lösung von Forschungsproblemen verwendet werden, die außerhalb der Bibliotheks- und Informationswissenschaft angesiedelt sind. Weil gerade wissenschaftliche Zeitschriftenliteratur für statistische Analysen gut geeignet ist und weil in der Wissenschaft nur das Publizierte zählt, kann Bibliometrie als wesentlicher Teil der Scientometrie auch bei der Beantwortung von Fragestellungen zu Struktur und Dynamik von Wissenschaft helfen. Die paradigmatische Arbeit von Lotka (1926) zur Verteilung der Produktivität wissenschaftlicher Autoren ging gerade von einer Frage aus, die man heute der Wissenschaftsforschung zuordnen würde. Lotkas Arbeit ist deshalb von heute aus betrachtet eine interdisziplinäre, bloß dass es Wissenschaftsforschung als Forschungsgebiet damals noch nicht gab und Bibliothekswissenschaft noch nicht bibliometrisch vorging.

Mehrere der im vorliegenden Text behandelten Themen sind für die Wissenschaftsforschung interessant: Neben den Verteilungen von Produktivität und Wirkung, den Koautorschaftsnetzwerken und Kozitationsclustern sind hier vor allem die bibliometrischen Modelle als Erklärungsversuche für die empirisch aufgefundenen schiefen Verteilungen zu nennen.

Alles das ist aber auch für die Bibliotheks- und Informationswissenschaft von Interesse, hilft es ihr doch, ihr wichtigstes Anwendungsgebiet, die Wissenschaft, besser zu verstehen. Sie profitiert damit also zweifach von bibliometrischen Analysen und Methoden, einmal direkt bei der Aufhellung von Gesetzmäßigkeiten von Informationsströmen, einmal indirekt über die Wissenschaftsforschung, deren Resultate für sie unverzichtbar sind.

Kapitel 7

Ausblick

Gutenbergs Erfindung des Buchdrucks mit beweglichen Lettern vor mehr als einem halben Jahrtausend ermöglichte es, nun sehr viel mehr und billiger Bücher zu produzieren. Dass Literatur dadurch quantitativ stark zunahm, führte zu qualitativen Umbrüchen in der Art und Weise, wie ihre Produktion, Verbreitung und Aufbewahrung organisiert wurde. Es entstand das über Jahrhunderte sich nur langsam verändernde System von Verlagen, Buchhändlern und Bibliotheken.

Heute erleben wir eine ähnliche Medienrevolution: Literatur, aber auch Bild- und Tonaufnahmen können jetzt in digitaler Form gespeichert und mit Lichtgeschwindigkeit über das Netz verbreitet werden. Ein Werk zu kopieren, kostet praktisch nichts mehr. Maschinelle Recherche im Volltext ist erstmals möglich. Am Ende werden zwar auch wissenschaftliche Zeitschriftenaufsätze immer noch zum Lesen ausgedruckt, aber nur weil sich das elektronische Papier als lesefreundliches (und energiesparendes) Ausgabemedium noch nicht durchgesetzt hat.

Dass wissenschaftliche und künstlerische Werke, Informationen aller Art nun so einfach verbreitet und kopiert werden können, lässt den Traum, allen Menschen umfassende Bildung zu ermöglichen, seiner Verwirklichung näher rücken. Der freie, d. h. kostenlose und unbeschränkte Zugang zu digitalen Kopien aller Werke, zu über das Netz verfügbaren Informationen schafft eine notwendige Voraussetzung der Teilhabe aller dazu überhaupt fähigen Menschen an Kultur, Politik und Wissenschaft, die mit Bibliotheken bisher nur teilweise verwirklicht werden konnte.

Die digitale und elektronische Informationstechnologie wird genauso die Produktions-, Verbreitungs- und Rezeptionsweise von Literatur verändern, wie es der Buchdruck Ende der ersten Hälfte des vorigen Jahrtausends tat. Verlage, Buchhändler und Bibliotheken müssen sich dieser Umwälzung anpassen oder werden untergehen. Wenn Verlage digital vorliegende Information künstlich verknappen, indem sie Kopiersperren erfinden oder – Gipfel des Widersinns – den Bibliotheken nur erlauben, jeweils *eine* Kopie *einem* Nutzer an *einem* Rechner in ihren Räumen lesen zu lassen, dann zeigt das, dass ihr Geschäftsmodell überlebt ist. Die Zukunft gehört *Open Access*, zumindest für die wissenschaftliche Zeitschriftenliteratur.

Heutige bibliometrische Forschung hat das Glück, die Medienrevolution analytisch begleiten zu können, aber nicht nur im Sinne einer quantitativen Erfassung der sich ändernden Verhältnisse, sondern auch durch das Aufzeigen von Entwicklungspfaden und die Aufhellung ihrer Voraussetzungen.

Sich änderndes Publikations- und Rezeptionsverhalten von wissenschaftlich Forschenden kann in seinen Zusammenhängen, seinen Tendenzen bibliometrisch und informetrisch untersucht werden. Einige Ansätze dazu liegen bereits vor (Gingras, Larivière, Archambault und Wray 2009).

Wissenschaft als methodisches Erzeugen neuen Wissens ist ihrem Wesen nach revolutionär. Ständig wird altes Wissen durch neues ersetzt. Ihr starkes exponentielles Wachstum im letzten Drittel des vorigen Jahrtausends hat auch das wissenschaftliche Kommunikationssystem einem ständigen Wandel unterworfen. Jetzt aber erleben wir zum Einem einen grundlegenden Umbruch dieses Systems und zum Anderen das Ende des exponentiellen Wachstums von Wissenschaft (Larivière, Archambault und Gingras 2008).

Nicht nur für die Bibliotheks- und Informationswissenschaft ist die heutige Zeit also interessant, auch für die Wissenschaftsforschung – und beide profitieren in ihren Analysen des Wandels von Bibliometrie.

Literatur

Ajiferuke, I., Q. Burrell und J. Tague (1988). Collaborative coefficient – a single measure of the degree of collaboration in research. *Scientometrics 14*, S. 421–433.

Barabasi, A. L. und R. Albert (1999). Emergence of scaling in random networks. *Science 286*, S. 509–512.

Bookstein, A. (1990). Informetric distributions, part II: Resilience to ambiguity. *Journal of the American Society for Information Science 41*(5), S. 376–386.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering 137*, S. 85–86.

Bradford, S. C. (1985). Sources of information on specific subjects. *J. Inf. Sci. 10*(4), S. 173–180. Reprint: Bradford (1934).

Brin, S. und L. Page (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems 30*(1–7), S. 107–117.

Burrell, Q. und R. Rousseau (1995). Fractional Counts for Authorship Attribution: A Numerical Study. *Journal of the American Society for Information Science and Technology 46*(2), S. 97–102.

Czerwon, H. und F. Havemann (1991). Deutsche physikalische Zeitschriften im Science Citation Index. *Physikalische Blätter 47*, S. 645–647.

de Solla Price, D. (1976). A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society of Information Science 27*, S. 292–306.

de Solla Price, D. J. (1963). *Little Science, Big Science*. New York: Columbia University Press. dt. mit selbem Titel bei Suhrkamp 1974.

de Solla Price, D. J. (1965). Networks of scientific papers. *Science 149*, S. 510–515.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian Informetrics*. Elsevier.

Egghe, L. und R. Rao (1992). Citation Age Data and the Obsolescence Function: Fits and Explanations. *Information Processing and Management 28*(2), S. 201–17.

Egghe, L. und R. Rousseau (1990). *Introduction to Informetrics. Quantitative Methods in Library and Information Science*. Elsevier.

Fermi, E. (1949). On the Origin of the Cosmic Radiation. *Physical Review 75*(8), S. 1169–1174.

Gabaix, X. (1999). Zipf's Law For Cities: An Explanation. *Quarterly Journal of Economics 114*(3), S. 739–767.

Garfield, E., A. Pudovkin und V. Istomin (2003). Why do we need algorithmic historiography? *Journal of the American Society for Information Science and Technology 54*(5), S. 400–412.

Garfield, E. und I. Sher (1963). New Factors in the Evaluation of Scientific Literature Through Citation Indexing. *American Documentation 14*(3), S. 195–201. `http://www.garfield.library.upenn.edu/essays/v6p492y1983.pdf`, 2005-4-28.

Geller, N. L. (1978). On the citation influence methodology of Pinski and Narin. *Information Processing & Management 14*(2), S. 93–95.

Gingras, Y., V. Larivière, É. Archambault und K. Wray (2009). Literature Citations in the Internet Era. *Science 323*, S. 36.

Glänzel, W. (2003). Bibliometrics as a research field: A course on theory and application of bibliometric indicators. Courses Handout, `http://www.norslis.net/2004/Bib_Module_KUL.pdf`.

Glänzel, W. (2008). Seven Myths in Bibliometrics. About facts and fiction in quantitative science studies. In: H. Kretschmer und F. Havemann (Hrsg.), *Proceedings of WIS 2008: Fourth International Conference on Webometrics, Informetrics and Scientometrics Ninth COLLNET Meeting*,

Berlin. Humboldt-Universität zu Berlin: Gesellschaft für Wissenschaftsforschung. http://www.collnet.de/Berlin-2008/GlanzelWIS2008smb.pdf.

Gross, P. und E. Gross (1927). College Libraries and Chemical Education. *Science 66*(1713), S. 385–389.

Havemann, F., M. Heinz und R. Wagner-Döbler (2005). Firm-like Behaviour of Journals? Scaling Properties of Their Output and Impact Growth Dynamics. *Journal of the American Society for Information Science and Technology 56*(1), S. 3–12. http://141.20.126.8/~fhavem/Havemann-Heinz-WagnerDoebler-Journals_Growth_Dynamics.pdf.

Hicks, D. und J. Katz (1996). Science policy for a highly collaborative science system. *Science and Public Policy 23*, S. 39–44.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences 102*(46), S. 16569–16572. http://arxiv.org/abs/physics/0508025.

Katz, J. und B. Martin (1997). What is research collaboration? *Research Policy 26*(1), S. 1–18.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation 14*, S. 10–25.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM 46*(5), S. 604–632.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press. Deutsch als *Die Struktur wissenschaftlicher Revolutionen* bei Suhrkamp: 1. Auflage 1973, 2. Auflage 1976.

Larivière, V., E. Archambault und Y. Gingras (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900-2004). *Journal of the American Society for Information Science and Technology 59*(2), S. 288–296.

Laudel, G. (1999). Interdisziplinäre Forschungskooperation: Erfolgsbedingungen der Institution 'Sonderforschungsbereich'. *Berlin, Edition Sigma*.

Leimkuhler, F. (1967). Bradford Distribution. *Journal of Documentation 23*(3), S. 197–207. s. a. Sonderdruck Purdue Univ. 1967.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences 16*(12), S. 317–323.

Magyar, G. (1974). Bibliometric Analysis of a New Research Sub-Field. *Journal of Documentation 30*(1), S. 32–40.

Marshakova, I. (1973). System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2 – Informatsionnye Protsessy i Sistemy 6*, S. 3–8. (in Russisch).

Marshakova, I. (1988). *Sistema tsitirovaniia nauchnoi literatury kak sredstvo slezheniia za razvitiem nauki (Scientific Literature Citation System as Means for the Science Evolution Monitoring)*. Moscow: Nauka.

Matricciani, E. (1991). The probability distribution of the age of references inengineering papers. *Professional Communication, IEEE Transactions on 34*(1), S. 7–12.

Mayr, P. (2008). An evaluation of Bradfordizing effects. In: H. Kretschmer und F. Havemann (Hrsg.), *Proceedings of WIS 2008: Fourth International Conference on Webometrics, Informetrics and Scientometrics Ninth COLLNET Meeting*, Berlin. Humboldt-Universität zu Berlin: Gesellschaft für Wissenschaftsforschung. http://www.collnet.de/Berlin-2008/MayrWIS2008ebe.pdf, Preprint arXiv:0812.0262.

Merton, R. (1968). The Matthew Effect in Science. *Science 159*(3810), S. 56–63.

Merton, R. K. (1988). The Matthew effect in science, II: cumulative advantage and the symbolism of intellectual property. *Isis 79*, S. 606–623.

Milgram, S. (1967). The small world problem. *Psychology Today 2*(1), S. 60–67.

Mitzenmacher, M. (2004). A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics 1*, S. 226–251. Retrieved from http://www.internetmathematics.org/volumes/1/2/pp226_251.pdf, date: 2004-7-27.

Moed, H. F., W. Glänzel und U. Schmoch (Hrsg.) (2004). *Handbook of Quantitative Science and Technology Research: The Use of Publication and Patent Statistics in Studies of S&T Systems*. Dordrecht: Kluwer Academic Publishers.

Nalimov, V. und Z. Mul'čenko (1969). *Naukometriâ*. Moskva: Nauka.

Naranan, S. (1970). Bradford's law of Bibliography of Science: an Interpretation. *Nature 227*, S. 631–632.

Naranan, S. (1971). Power law relations in science bibliography – self-consistent interpretation. *Journal of Documentation 27*, S. 83–97.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review 45*, S. 167–256.

Newman, M. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics 46*(5), S. 323–351.

Newman, M. E. J. (2001a). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E 64*, S. 016131.

Newman, M. E. J. (2001b). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E 64*, S. 016132.

Oluic-Vukovic, V. (1997). Bradford's distribution: From the classical bibliometric 'law' to the more general stochastic models. *Journal of the American Society for Information Science 48*(9), S. 833–842.

Persson, O., W. Glänzel und R. Danell (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics 60*(3), S. 421–432.

Pinski, G. und F. Narin (1976). Citation influence for journal aggregates of scientific publications: theory, with application to literature of physics. *Information Processing & Management 12*, S. 297–312.

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation 25*(4), S. 348–349.

Radicchi, F., C. Castellano, F. Cecconi, V. Loreto und D. Parisi (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America 101*, S. 2658–2663.

Rescher, N. (1978). *Scientific progress: A philosophical essay on the economics of research in natural science*. Oxford.

Rescher, N. (1982). *Wissenschaftlicher Fortschritt: Eine Studie über die Ökonomie der Forschung*. Walter de Gruyter. Dt. Fassung des Buches von Rescher (1978).

Rousseau, R. (1988). Lotka's law and its Leimkuhler representation. *Library science with a slant to documentation and information studies 25*(3), S. 150–178.

Rousseau, R. (2002). George Kingsley Zipf: life, ideas, his law and informetrics. *Glottometrics 3*, S. 11–18.

Salton, G. und M. J. McGill (1983). *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc.

Salton, G. und M. J. McGill (1987). *Information Retrieval – Grundlegendes für Informationswissenschaftler*. McGraw-Hill.

Schubert, A. und T. Braun (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics 9*(5), S. 281–291.

Schubert, A., W. Glänzel und B. Thijs (2006). The weight of author self-citations. A fractional approach to self-citation counting. *Scientometrics 67*(3), S. 503–514.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science 43*, S. 628–638.

Shockley, W. (1957). On the Statistics of Individual Variations of Productivity in Research Laboratories. *Proceedings of the IRE 45*(3), S. 279–290.

Simon, H. (1955). On a Class of Skew Distribution Functions. *Biometrika 42*(3/4), S. 425–440.

Small, H. (1973). Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science 24*, S. 265–269.

Small, H. und E. Sweeney (1985). Clustering the Science Citation index® using cocitations. *Scientometrics 7*(3), S. 391–409.

van Raan, A. (1990). Fractal dimension of cocitations. *Nature 347*, S. 626.

van Raan, A. (2000). On growth, ageing, and fractal differentiation of science. *Scientometrics 47*(2), S. 347–362.

Vinkler, P. (2000). Evaluation of the publication activity of research teams by means of scientometric indicators. *Current Science 79*(5), S. 602–612.

Wasserman, S. und K. Faust (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

White, H. (1981). 'Bradfordizing' search output: how it would help online users. *Online Information Review 5*(1), S. 47–54.

Zipf, G. (1932). *Selected studies of the principle of relative frequency in language*. Harvard University Press, Cambridge, Mass.

Zipf, G. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Biology*. Cambridge, Mass.: MIT Press.

Zipf, G. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press.