UNIFIED MATHMATICAL TREATMENT OF COMPLEX CASCADED BIPARTITE NETWORKS: THE CASE OF COLLECTIONS OF JOURNAL PAPERS

By

STEVEN ALLEN MORRIS

Bachelor of Science
Tulsa University
Tulsa, Oklahoma
1983

Master of Science
Tulsa University
Tulsa, Oklahoma
1987

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2005

UNIFIED MATHMATICAL TREATMENT OF
COMPLEX CASCADED BIPARTITE
NETWORKS: THE CASE OF
COLLECTIONS OF
JOURNAL PAPERS

Thesis Approved:

Dr. Gary Yen
Thesis Advisor

Dr. Chris Hutchens

Dr. Martin Hagan

Dr. David Pratt

Dr. A. Gordon Emslie
Dean of the Graduate College

PREFACE

In this study, a mathematical treatment is proposed for analysis of entities and relations among entities in complex networks consisting of cascaded bipartite networks. This treatment is applied to the case of collections of journal papers. In this case, entities are distinguishable objects and concepts, such as papers, references, paper authors, reference authors, paper journals, reference journals, institutions, terms, and term definitions. Relations are associations between entity-types such as papers and the references they cite, or paper authors and the papers they write. An entity-relationship model is introduced that explicitly shows direct links between entity-types and possible useful indirect relations. From this a matrix formulation and generalized matrix arithmetic are introduced that allow easy expression of relations between entities and calculation of weights of indirect links and co-occurrence links. Occurrence matrices, equivalence matrices, membership matrices and co-occurrence matrices are described. A dynamic model of growth describes recursive relations in occurrence and co-occurrence matrices as papers are added to the paper collection. Graph theoretic matrices are introduced to allow information flow studies of networks of papers linked by their citations. Similarity calculations and similarity fusion are explained. Derivation of feature vectors for pattern recognition techniques is presented. The relation of the proposed mathematical treatment to seriation, clustering, multidimensional scaling, and visualization techniques is discussed. It is shown that most existing bibliometric analysis techniques for dealing with collections of journal papers are easily expressed in terms of the proposed mathematical treatment: co-citation analysis, bibliographic coupling analysis, author co-citation analysis, journal co-citation analysis, Braam-Moed-vanRaan (BMV) co-citation/co-word analysis, latent semantic analysis, hubs and authorities, and multidimensional scaling. This report discusses an extensive software toolkit that was developed for this research for analyzing and visualizing entities and links in a collection of journal papers. Additionally, an extensive case study is presented, analyzing and visualizing 60 years of anthrax research through a collection of journal papers. When dealing with complex networks that consist of cascaded bipartite networks, the treatment presented here provides a general mathematical framework for all aspects of analysis of static network structure and network dynamic growth. As such, it provides a basic paradigm for thinking about and modeling such networks: computing direct and indirect links, expressing and analyzing statistical distributions of network characteristics, describing network growth, deriving feature vectors, clustering, and visualizing network structure and growth.

ACKNOWLEDGEMENTS

I wish to express my appreciation to my major advisor, Dr. Gary Yen, for his guidance and encouragement. I appreciate the freedom and trust that he gave me during my studies and research. I am grateful to Dr. Chris Hutchens for encouraging me to start this whole journey, and for providing financial support when it was needed. I also deeply appreciate the help of Dr. Tom Collins, who encouraged the original research and helped with financial support by employing me on the ASSET project. It was a great pleasure to have fellow student Michel Goldstein as a coworker, his ideas on ontology research were the inspiration for the entity-relation model that is the basis of this study.

Finally, I wish to express my heartfelt appreciation to my wife Esther who never faltered in her support throughout the whole project. Both Esther and my daughter Jenny showed great patience and cheerfulness throughout the whole journey.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

1. INTRODUCTION

1.1 Motivation

Collections of journal papers are defined as databases of information about journal papers whose subjects broadly focus on some scientific specialty. The papers and the entities associated with them, that is, paper authors, paper journals, references, reference authors, reference journals, and terms, form complex networks whose underlying structure is a manifestation of the structure of the scientific specialty, its research sub-topics, paradigm, exemplars, invisible colleges, collaboration groups, and archiving journals.

It is the goal of this study to introduce a unified mathematical treatment of the entities and entity links that comprise a collection of papers. This mathematical treatment will serve to codify many existing concepts that pertain to journal paper collections, and greatly simplifies and consolidates the understanding and application of many bibliometric analysis techniques. In this sense, the proposed mathematical treatment to be introduced here is a unified mathematical model of the networks that comprise collections of journal papers.

The existence of simple mathematical models of the concepts and principles within a scientific specialty is tremendously useful to researchers within the specialty. As an example, in the field of electrical engineering there is no principle as ubiquitous as Ohm's Law. This principle is always spelled out in equation form: E = IR, signifying that "the voltage across a resistor is equal to its resistance times the value of the electric current through that resistor." This principle is so simple, so basic, and seemingly so obvious, that electrical engineers use it with absolutely no doubt of its validity.

Yet, the very name 'Ohm's Law' implies that the principle was discovered, and prior to that discovery, researchers struggled along in the dark without it. Interestingly, as Kuhn (1970, p. 183) pointed out, Ohm's Law, at its discovery, not only defined the mathematical relation between electrical current, voltage and resistance, it defined the very concepts of electrical current and resistance themselves (Schagrin, 1963), concepts without which electrical engineering would be very confused indeed. Considering its supposed obviousness, it is surprising that Ohm's Law was, in fact, accepted only after considerable resistance (Schagrin, 1963).

The example of Ohm's Law points to the importance and necessity of a mathematical treatment of a phenomenon to be studied. Indeed it would appear that the ability to mathematically describe their problems is what separates the physical scientists from the social scientists and allows the formation of paradigms wherein Kuhnian puzzle-solving, the efficient mode of 'normal science' (Kuhn, 1970, p. 178), can occur.

A complete and consistent mathematical formulation of a phenomenon under study within a specialty has two desirable benefits: 1) it standardizes the problem in a way that enables practitioners within the specialty to efficiently communicate about the problems they are working on , and 2) it enables researchers to view with great insight the structures and symmetries of the subject that they are studying. As Crane reports (1980), scientists judge proposed mathematical models not only for their potential use and testability, but also for their 'elegance.' Elegance is a subjective term that describes a model that is simple, concise, starkly symmetric, and insightful. As an example, for a practicing electrical engineer, the use of complex arithmetic to analyze phase relationships in circuits is not only wonderfully useful, but aesthetically satisfying and elegant, changing a complicated process of muddling through error-prone trigonometric calculations, to the simple addition and subtraction of real and imaginary parts of complex numbers[1].

There is little unified mathematical treatment in the study of bibliometrics. There have been a number of mathematical models of the process that generates the ubiquitous power law distributions that always appear when studying collections of journal papers. In fact, it seems as if a whole specialty in bibliometrics is built on studying a series of eponymous empirically observed mathematical distributions: Lotka's Law, Bradford's Law, Zipf's Law, and so forth (White & McCain, 1989). Of course, many researchers have noticed the similarities among these various 'laws' (Bookstein, 1990; Fairthorne, 1969; Seglen, 1992). Nevertheless, there has been no study of these distributions and how they are related based on a unified mathematical treatment of the entities in collections of journal papers.

Another rich area of study in bibliometrics has been the study of the process of citation of references (White & McCain, 1989). These models describe stochastic processes that generate distributions of citations to references and *literature aging*, the study of citation rates to references over time. Many of these models are mathematically sophisticated, but all of them stand alone, that is, they are not based on any unified mathematical description of the entities in a collection of journal papers.

This report will present a mathematical treatment of collections of journal papers. The term 'mathematical treatment' is used rather than 'mathematical model' to distinguish this treatment from models of underlying

[1]While physicists and mathematician often speak of elegance with admiration, it is worth remembering Albert Einstein's negative comment that "elegance is for tailors and cobblers."

processes in a research specialty and the manifestation of those processes in the specialty's literature. While it is intended that the mathematical treatment presented here will facilitate construction of unified models of research processes, the consideration of such processes is beyond the scope of the discussion in this report. Nonetheless, the proposed mathematical treatment, based on matrix arithmetic, is concise, consistent, insightful, and, if an investigator is using a scientific software package designed for matrix manipulation and analysis, quite effective in practice. This mathematical treatment can be used to represent the entities and the links among entities in collections of papers:

- **Bibliometric entities and entity-types**. Definition of bibliometric entities helps define what is being studied in bibliometric analysis and further allows systematic study of the manifestation of the progress of research in a scientific specialty in the literature. Examples of entities are papers, references, and authors of papers.
- **Direct links among entities**. Direct links are direct, observable associations between pairs of entities in the collection of journal papers. They are manifested directly as lists of associations implemented as tables in the database of the collection of papers. For example, papers are associated with the authors that wrote them, the references they cite, the journals they appear in, and the terms they use. Reference authors are associated with one or more references. These associations constitute direct links between entities in the collection of papers.
- **Indirect links among entities**. Indirect links are associations between pairs of entities that are found by calculating paths through networks of direct links. Examples of indirect links are: links from paper authors to reference authors, or links from index terms to references. The mathematical treatment introduced here will present a general method for calculating indirect links between entities.
- **Co-occurrence links among entities**. Co-occurrence links between like entities occur when two like entities are associated with the same unlike entity. As examples, two paper authors are linked when they co-author a paper, or two papers are linked when they both cite the same reference, or two references are linked if they both are cited in the same paper. Co-occurrence links are the main tool used for mapping the structure of a research specialty through its manifestation in the scientific literature.

Using the mathematical representation of entities and the links among entities as described in the list above, many techniques can be used to manipulate and analyze these entities for bibliometric analysis. Most bibliometric analysis focuses on analysis of network link structure and this analysis can be divided into four main types of applications:

1) **Ranking and evaluation**. The ranking of the importance and impact of researchers, journals, and institutions is commonly used to provide a tool for assessment of research performance. Most of

the methods of ranking are based on derivation of impact factors and research indicators based on patterns of citations in a specialty or field (White & McCain, 1989).

2) **Structural mapping**. Mapping focuses on building abstract models of the network of knowledge in a specialty and the social structure of scientists that participate in research in that specialty (Borner, Chen, & Boyack, 2002). This is the process of finding and labeling meaningful entity groups and the relations between them.
3) **Search**. Search focuses on finding key pieces of knowledge, critical avenues of communication, or important entities within the scientific specialty. This is the process of finding important *individual* entities and relations (Salton, 1989). This contrasts with structural mapping, which is concerned with finding *groups* of entities.
4) **Forecasting.** This focuses on the dynamic changes occurring within a specialty and attempts to extrapolate those changes into the future. Forecasting deals with both analysis of trends and detection of discontinuous events (Morris, DeYong, Wu, Salman, & Yemenu, 2002; Zhu & Porter, 2002).

As will be seen in the remainder of this report, an *entity-relationship (ER) model* of the collection of papers, and a *matrix formulation* of the relations among entities is the basis for a mathematical treatment that is introduced here. This treatment is a tool that facilitates the four types of link structure analysis discussed above.

A secondary application of bibliometrics is in *concept evolution studies*. In a historical sense, concept evolution attempts to trace the evolution of ideas and flow of knowledge leading up to and following a scientific discovery (Garfield, Pudovkin, & Istomin, 2003). The emphasis in this type of analysis is on *information flow* from one entity to the next. For this type of analysis a *graph theoretic model* of collections of papers is necessary. For example, it is useful to present papers in the collection as a graph, where each paper is a node and the links between them are citations. As will be seen later, the necessary graphs for this type of analysis can be easily derived from the entity-relationship model.

The goals of this study are: 1) to produce a consistent notation and mathematical treatment that will allow a researcher to obtain a simple view of the data and analysis techniques associated with paper collections. 2) to suggest preferred methods and tools for analysis according to the applications and, 3) to consolidate existing bibliometric analysis techniques within the framework of a unified mathematical treatment.

1.2 Summary

The remainder of this report is organized as follows.

- In Chapter 2, an entity-relationship (ER) model of collections of journal papers is introduced. Entities and relationships are defined, the entities within a collection are enumerated through an entity-relationship diagram, and a mathematical notation is introduced to represent entities and relations.
- In Chapter 3, a generalized matrix arithmetic is introduced that is used for calculation of link weights and indirect links.
- In Chapter 4, the concept of occurrence matrices is introduced to express relations between entities. The calculation of occurrence matrices of indirect links is demonstrated. Equivalence matrices are introduced to relate pairs of bibliometric entities to physical entities, while membership matrices are introduced for calculating group memberships of entities.
- In Chapter 5, co-occurrence matrices are introduced. The calculation of co-occurrence matrices is explained using matrix multiplication for calculation from binary occurrence matrices and using the overlap function and other generalized link weight functions for calculation of co-occurrence matrices from non-binary occurrence matrices.
- In Chapter 6, bibliometric distributions are discussed. This includes an explanation of static and cumulating dyadic occurrence distributions as well as the introduction of co-occurrence distributions and clustering coefficient distributions.
- In Chapter 7, a recursive matrix formulation is introduced to express the growth of a collection of papers. This recursive formulation is used to show the growth of occurrence matrices as well as the growth of static and cumulating co-occurrence matrices.
- In Chapter 8, a formulation of graph theoretic methods in bibliometrics is introduced. The calculation of graph theoretic linkages from occurrence matrices and equivalence matrices is explained.
- In Chapter 9, the calculation of inter-entity similarities in terms of the proposed mathematical treatment is introduced. Fusion of multiple similarities is introduced and these fusion methods are used to express Small's similarity that is based on graph theoretic methods.
- In Chapter 10, the concept of entity feature vectors is introduced, and feature vectors based on both occurrence and co-occurrence are explained.
- In Chapter 11, the relation of the proposed mathematical treatment to seriation and clustering is explained. Seriation, matrix shading, agglomerative hierarchical clustering, and c-means clustering are discussed.
- In Chapter 12, visualization of occurrence matrices is explained. Timelines, usage plots, and crossmaps are introduced.
- In Chapter 13, existing bibliometric analysis techniques are reviewed in terms of the proposed mathematical formulation. These techniques include:

    - Co-citation analysis

- Bibliometric coupling analysis
- Author co-citation analysis
- Journal co-citation analysis
- Braam-Moed-vanRaan (BMV) co-citation/co-term analysis
- Latent semantic analysis
- Hubs and authorities analysis.
- White's concepts of author identities and author images
- Pathfinder analysis

- In Chapter 14, a software toolkit for analyzing and visualizing collections of journal papers using the methods developed through this research is presented.
- In Chapter 15, a case study on the specialty of anthrax research is presented, which illustrates the application of the methods developed through this research.
- In Chapter 16, there is a concluding discussion of the possible impact of the proposed mathematical formulation on bibliometric theory and analysis. Notes on more software implementations of analysis techniques are discussed. Directions for continued research to extend the proposed mathematical treatment also discussed.

2. ENTITY-RELATIONSHIP MODEL OF A COLLECTION OF JOURNAL PAPERS

2.1 Definition of collections of journal papers

Collections of journal papers, as discussed in this report, are databases of information about journal papers. In the database are lists of papers and the data associated with each paper: title, abstract text, published journal information, references, paper authors, and index terms. For the discussion here, it is assumed that the body text of the paper is not stored in the database. For brevity, from this point forward, collections of journal papers will be referred to as 'collections' or 'collections of papers.'

These collections typically consist of papers that broadly cover the topic of some scientific specialty. The papers and the entities associated with them, paper authors, paper journals, references, reference authors, reference journals, and terms, form complex networks whose underlying structure mirrors the structure of the scientific specialty, its research sub-topics, exemplars, invisible colleges, collaboration groups, and archiving journals. Collections of papers are typically gathered from an abstracting service, very often from the Science Citation Index, but also from other services such as Chemical Abstracts or Petroleum Abstracts.

The papers are usually gathered by using queries against index terms, by using seed references (by finding all papers that cite some set of base references in the specialty), or by using seed reference authors (by finding all papers that cite some important reference author's oeuvre.) Databases holding collections of journal papers typically do not contain the papers' body text and figures, and may or may not contain lists of references cited by each paper. For discussion purposes, this study assumes that the body text of papers is not available and that the references cited by papers are available.

2.2 Entities in collection of journal papers

Using an entity-relationship (ER) model, a collection of journal papers can be considered as a collection of entities of different entity-types (Morris & Yen, 2004). Define *bibliometric entities* as objects of interest in the collection of papers. Examples of bibliometric entities include papers, references, and terms. The objects from the physical world, to which bibliometric entities may correspond, are defined as *physical entities*. Bibliometric entities are manifestations of *physical entities*, and a single physical entity may correspond to more than one bibliometric entity. For example, paper author "H. Small", and the two

reference authors "Small H" and "Small HG" are three bibliometric entities that all correspond to the same physical person, Henry Small. In another example, a reference identified as "SMALL-H-1978-SOC-STUD-SCI-V8-P327" and a paper whose title is "Cited documents as concept symbols" both correspond to a physical paper written by Henry Small and published in *Social Studies of Science* in 1978.

In much of bibliometric analysis, it is quite useful to work with a bibliometric entity without regard to the physical entity to which it corresponds, and without regard to other bibliometric entities that correspond to the same physical entity. As will be explained in the next paragraph, cited bibliographic entities pick up symbolic meaning that is detached from the original physical entity. It is, however, often necessary to equate pairs of bibliometric entities through their corresponding physical entities. This procedure is necessary, for example, when using paper collections to trace concept evolution, where, for the purpose of tracing information flow, it is necessary to equate cited entities to citing entities. In this report, the term "entity" or "entities" will refer to bibliometric entities unless specifically stated otherwise.

The great power of studying entities within a collection of papers is in the analysis of those entities as symbols representing *research elements*. In this report, research elements are defined as the objects that comprise scientific research, such as researchers, research topics, institutions, invisible colleges, funding sources, and corporate and government research consumers (Leydesdorff, 1995). Taking this definition a step further, the *Kuhnian research elements* are paradigms, exemplars, and puzzles (Kuhn, 1970). The goal of bibliometric analysis is to find these abstract objects and groups of objects and analyze the relations among them. The following entity-types and their related symbolic meanings are of interest:

- **Papers**. These are *research reports* and comprise the basic unit within a collection of papers. The collection of papers grows one paper at a time.
- **References**. In this study references are objects that are *cited* by papers. References are objects while citations are actions. References can, in a general sense, be considered as *concept symbols*, especially those references that are heavily cited.
- **Paper journals**. These are the journals within which papers are published. In general, paper journals can be considered as *research report archives* for the specialty.
- **Paper authors**. Paper authors are the creators of papers and can be considered as *researchers*.
- **Reference journals**. These are journals which are associated with references. They can be considered *base knowledge archives* for a specialty. For a specialty which borrows much *loan knowledge* from other fields, there may be little overlap of reference journals and paper journals.
- **Reference authors**. These are authors associated with references. In a general sense, they can be considered as *base knowledge generators*, or *experts*, but can also be considered as symbols of *schools of thought*. Heavily cited authors are often studied in order to map the knowledge structure of the specialty (White & Griffith, 1981) .

- **Terms**. These are words or phrases that are *concept symbols*. Many types of terms can be of interest. *Author keywords*, with little quality control, are supplied by paper authors for searching and classification purposes. *Index terms*, are supplied by journals or abstracting services and are often taken from a fixed thesaurus, generating a fixed pool of consistent, easily searched terms. Keywords and index terms are associated only once per paper. *Linguistic terms* are terms extracted from prose within titles, abstracts, and the main text body of papers. *Title terms* are extracted automatically from titles of papers by software that parses titles and extracts single and multiword terms. A stop word list is usually used to eliminate useless terms. *Abstract terms* are similar to title terms but are taken from paper abstracts. Title and abstract terms can have multiple associations with papers they appear in.
- **Paper author institutions**. These can be considered as *research generating organizations*. In practice, it is difficult to use this information because of the need to extract consistent institution names from author addresses, which tend to contain ambiguously and inconsistently used institution titles, institution sub-division titles, and a great array of confusing institution acronyms.

The entity-types listed above form a general list of entity-types that are commonly studied. Several auxiliary entity-types can also be studied:

- **Paper publication year**. Although this is often treated as an attribute of papers, it is possible to define year as an entity-type for knowledge mapping purposes. Such mapping can be used to map trends and events in research activity.
- **Reference year**. Similar to paper publication years, this is often treated as an attribute of references. Reference year has been used for literature obsolescence studies (Burrell, 2001). However, as concept tokens, interpretation of reference age may have some interpretational problems. For example, a study of exemplar references in emerging specialties (Morris, 2004) indicates that it may be useful to consider reference age as the time since its first use in the specialty's literature rather than the time since the publication of the book or paper corresponding to the reference.
- **Paper country**. This is an entity which is easier to acquire from paper author addresses than author institutions. It can be used to study research activity by country.
- **Reference country**. This type of entity is more difficult to acquire than paper country because it must be found by matching reference authors to paper authors. Once acquired, reference country can be used to study production and flow of base knowledge among countries.

2.3 Relationships among entities in collections of journal papers

Within a collection of journal papers the entities are associated with each other. For example, a paper is associated with the paper authors who wrote it, the references it contains, the journal it was published in, and the terms that it contains. For the purposes of the mathematical treatment to be proposed here, all the direct links among entities are always defined between two differing entity-types. There are no direct associations between entities of the same entity-type. At first glance, this seems odd, because intuitively it appears that papers should be directly linked to the papers that they cite.

There are two reasons why citation links between papers must be considered as indirect links. First, consider that data from ISI's Web of Science product, which is a typical source of collections of papers, is supplied paper by paper and contains lists of references contained in each paper. There is no table provided to relate references to the actual papers to which they correspond. There are many references that correspond to books, reports, and web pages, for which there are no corresponding papers. Furthermore, since there are typically an order of magnitude more references than papers in a collection of papers, 90% of the references in the collection have no corresponding papers in the collection. Not only this, highly cited references typically have several versions in a paper collection, brought about by misspellings, missing information, and transcribing errors of authors and abstract services.

Secondly, even without the problem of matching references to papers, it would be necessary to treat them as separate entity-types because of the difference in symbolic meaning of references and papers. References are concept symbols while papers are research reports. The fact that a reference may correspond to a paper isn't really relevant to the meaning that the reference acquires as a concept symbol. It is therefore necessary to separate papers and references as separate entities. Similarly, reference authors and paper authors are semantically separate entity-types, and reference journals and paper journals are semantically different entity-types.

In the literature on entity-relationship models (Chen, 1976) it is common to differentiate between 'one-to-one,' 'one-to-many,' and 'many-to-many' relationships. For the ER model of collections of papers, it is also necessary to deal with the question of uniqueness in the associations. Uniqueness is important when dealing with questions of calculating link weights for co-occurrence counts and indirect links. Another concept of importance is the distinction between *independent entities* which appear and exist on their own and *dependent entities* which can only exist in association with an independent entity. In the ER model of paper collections, papers are independent, references are dependent on papers, and reference authors and reference journals are dependent on references. These dependencies are important for considering growth dynamics of collections of papers. Also in the ER modeling methods, it is common to list the relation between entities explicitly. For example, if there are two entity-types, employers and employees, then an

employer entity 'employs' employee entities, and employee entities 'work for' employer entities. Starting with papers, 6 pairs of direct relations are possible in collections of papers:

- **Papers-paper authors**: paper authors appear once in multiple papers, papers contain multiple unique paper authors.
- **Papers-paper journals**: paper journals contain multiple unique papers; papers appear in one journal once.
- **Papers-references**: references appear once in multiple papers; papers contain multiple, unique references.
- **Papers-terms**: 1) title and abstract terms appear multiple times in multiple papers; papers contain multiple terms multiple times. 2) index terms appear once in multiple papers; papers contain multiple unique index terms.
- **References-reference authors**: reference authors appear once in multiple references, references contain one reference author.
- **References-reference journals**: references journals appear once in multiple references; references contain one reference journal.
- **Paper author-institution**: Institutions contain multiple paper authors.

2.4 An entity-relationship diagram for collections of journal papers

Figure 1 shows an entity-relationship (ER) diagram of a collection of journal papers. The entity-types are denoted as circles while lines denote direct relationship pairs. Each relation should be read from the base of the arrow to the tip of the arrow. For example, the arrow going from papers to paper authors should be read "a paper contains multiple unique paper authors."

Note the central position in this diagram played by papers. As noted previously, papers are the basic unit in a collection of papers, data is added to the collection one paper at a time, with associated dependent entities and attributes. Papers are directly associated with: 1) the paper authors who write them, 2) the references that the papers cite, 3) the paper journals that they appear in, and 4) the terms that appear in them. In the database for the collection, database tables corresponding to each of these four dependent entity-types will typically exist, with each entity indexed by the paper that it is associated with. The lines between entity-types shown in Figure 1 represent *direct links*, that is, direct associations as discussed in Chapter 2.3.

Indirect links between two entity-types on the diagram are formed by chaining direct links on the diagram. For example, it is possible to find indirect links from paper authors to reference authors by starting with the paper authors, finding the papers they are associated with, finding the references cited by those papers, then finding the reference authors associated with those references.

A second entity-relationship diagram is possible, as shown in Figure 2, if the correspondences of bibliometric entities to physical entities are considered. Note that for clarity, in this figure the terms and institutions have been left off the diagram. Here it is possible to see multiple types of indirect links between entities, and indirect links that are circular, that is, indirect links between like entities. For example, it is possible to find links from paper authors to paper authors by starting with paper authors, finding the papers associated with them, finding the references those papers cite, finding the reference authors associated with those references, finding the physical authors associated with those reference authors, then finding the paper authors that correspond to those physical authors.

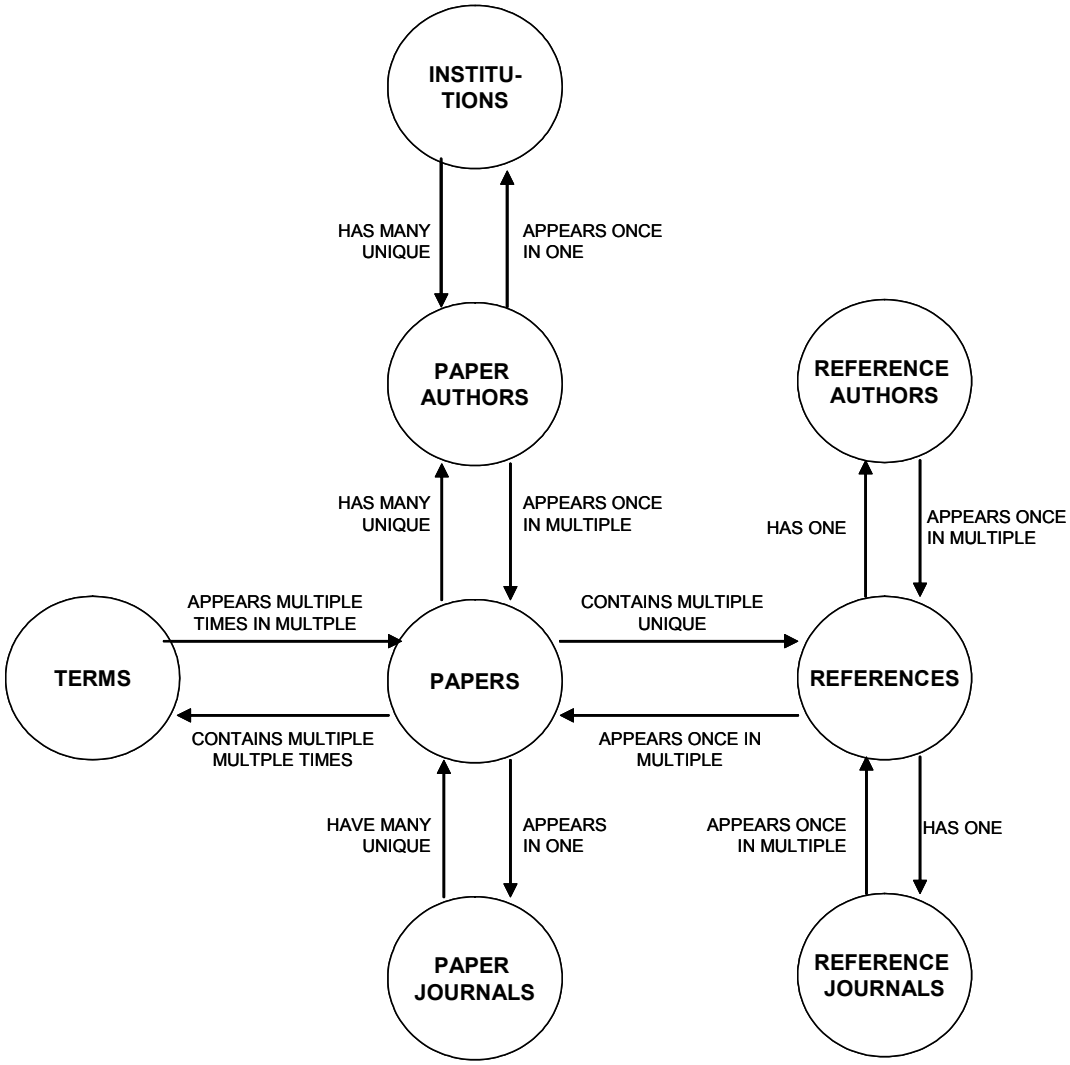


Figure 1. Entity-relationship diagram of a collection of journal papers.

2.5 Dyadic links and dyad notation

Table 1 lists the conventions used in this report to denote entity-type variables within the collection of papers. The variables x_1, x_2, and so forth will be used to denote unspecified entity-types. Specific entity-

types are denoted by the first letter of the name of the entity-type, e.g., *p* denotes paper, *r* denotes reference. Paper authors and other entity-types that have two words in their entity-type names use two letter variables with the letter for the dependent entity-type first. For example, *ap* for paper author. Additionally, to denote the number of entities of a particular entity-type in the collection of papers, attach an *n* ahead of the entity variable listed above, e.g., *np* is the number of papers in the collection, while *nr* is the number of references in the collection. Attach a *g* ahead of an entity-type variable to denote groups of entities of that entity-type. For example, *gp* denotes groups of papers.

In this report all links are dyadic, that is, they occur between two entities at a time. In a dyadic link, the two entities can be: 1) *like entities*, that is, entities of the same entity-type, or 2) *unlike entities*, that is, entities of different entity-type. *Occurrence links* are between unlike entities that are associated with each other. For example, there is an occurrence link between a paper and reference if the paper is associated with the reference by having cited it. The first entity of interest in a dyad is the *primary entity* while the other entity is the *relative entity*. *Co-occurrence links* are between like entities and occur if the like entities of the dyad are both associated with one or more common unlike entities. For example, two papers are linked when they both reference one or more identical references, or two paper authors are linked if they co-author one or more papers. In co-occurrence links the like entities of the dyad are primary entities, while the unlike entities with which they co-occur are the relative entities.

Table 1. Variable conventions used in this report for entities in collections of journal papers.

• *p* – paper	• *jr* – reference journal
• *r* – reference	• *jp* – paper journal
• *cp* – paper as reference	• *yp* – paper year
• *ap* – paper author	• *yr* – reference year
• *ar* – reference author	• *ip* – paper institution
• *t* - term	

Note:

- prefix *n* to any entity variable to denote the number of entities in the collection of that entity-type, e.g., *np* denotes the number of papers in the collection
- prefix *g* to any entity variable to denote groups of entities of that entity-type, e.g., *gr* denotes groups of references.

Dyad identifier notation. For notation in this report, the symbols of primary and relative entity-types associated with dyads will be separated by a semicolon and placed between square braces: $[x_1;x_2]$, where x_1 denotes the primary entity-type, and x_2 denotes the relative entity-type. This notation will be referred to as the *dyad identifier*. This notation will be used as a suffix to variables and in functions as well to specify the

entity-types of interest. However, the dyad identifier will be dropped to reduce clutter in the notation when the primary and relative entity-types are obvious from context. Some examples of the use of dyad identifiers:

- **O**[*p;r*] denotes the occurrence matrix listing the links of papers, the primary entity-type, to references, the relative entity-type.
- **C**[*ap;p*] denotes the co-occurrence matrix listing the co-authorship counts of paper authors, the primary entity-type, in papers, the relative entity-type.
- $p(k, \mathbf{O}[ap;p])$ denotes the "paper per paper author" distribution, that is, the probability that a paper author, the primary entity-type, will be associated with k papers, the relative entity-type.

The ER model of collections of papers, the concepts of direct and indirect links, bibliometric and physical entities, and the notation introduced in this chapter provide the basic framework for the mathematical treatment that will be introduced. Before introduction of the mathematical treatment it will be necessary to use Chapter 3 to discuss the method to compute indirect links from direct links in collections of papers.

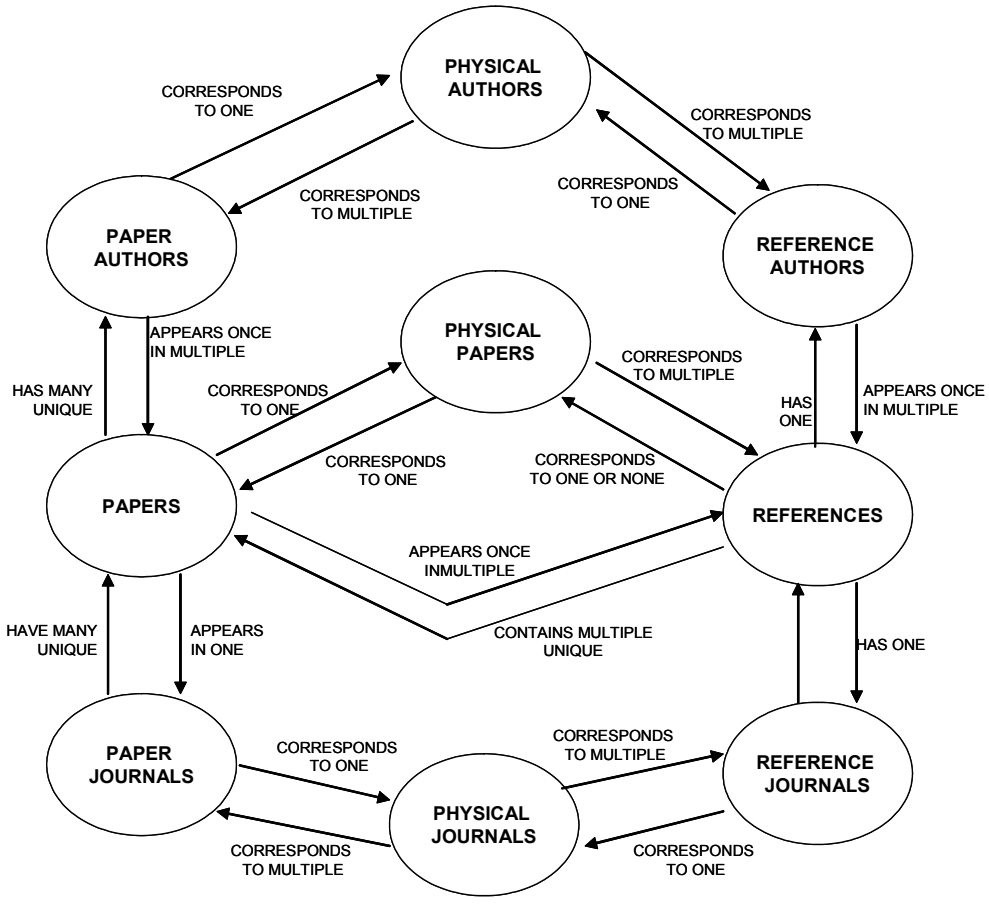


Figure 2. Entity-relationship diagram of a collection of journal papers showing links to physical entities by citing and cited bibliometric entities.

3. COMPUTATION OF LINK WEIGHTS

3.1 Bipartite networks

Bipartite networks are graphs comprised of two distinct groups of nodes, where all links in the graph are from entities in the first group to entities in the second group. As an example, Figure 3 shows a diagram of a bipartite graph of a group of papers linked to a group of references. Note that there are no links between papers or links between references.

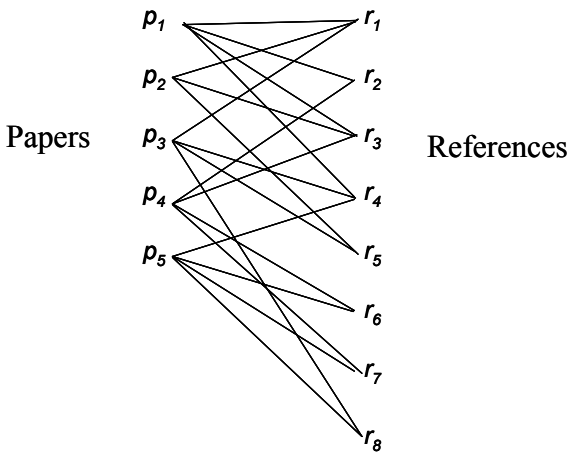


Figure 3. A collection of papers and references as a bipartite graph. References are linked to papers in which they are cited.

A bipartite network is comprised of entities of unlike entity-types. Assume the diagrammatic convention as shown in Figure 4, that entities of x_1, the primary entity-type, are the entities in the group on the left and the entities of x_2, the relative entity-type, are the entities in the group to the right. There are nx_1 primary entities and nx_2 relative entities. The strength of the link between x_1 entity i and x_2 entity j is the *link weight*, $o_{ij}[x_1;x_2]$. The following conventions are used in this report for link weights:

1. The magnitude of the link weight is proportional to the strength of the connection between two nodes.
2. Nodes with no connection have zero link weight.
3. Link weights can range from zero to positive infinity.

Mathematically, the links in the bipartite network are described by an *occurrence matrix*, analogous to an adjacency matrix in graph theory. The occurrence matrix is an nx_1 by nx_2 matrix that lists all the link weights between the entities of the two unlike entity-types:

$$\mathbf{O}[x_1;x_2]=\begin{bmatrix} o_{11} & o_{12} & \cdots & o_{1nx_2} \\ o_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ o_{nx_1 1} & \cdots & \cdots & o_{nx_1 nx_2} \end{bmatrix} \tag{1}$$

There is a bipartite network for every possible pair of unlike entity-types in the collection of papers. Given *NE* entity-types in the collection, there are *NE*(*NE*-1)/2 bipartite networks in the collection. Occurrence matrices for entity-type pairs with direct relations are derived directly from the tables in the collection's database. Occurrence matrices for entity-type pairs with indirect links are calculated by cascading bipartite networks of direct links, as will be shown later.

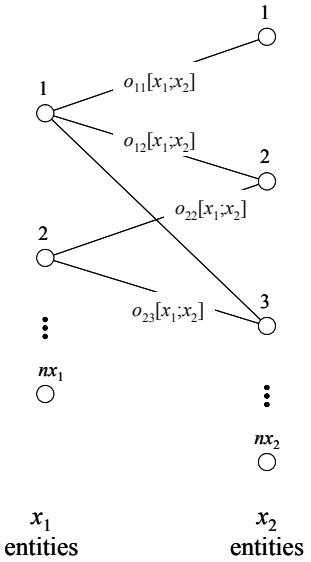


Figure 4. Diagram of a general bipartite network and conventions for naming entities and links.

3.2 Cascaded bipartite networks

Many of the entity-type pairs of interest in collections of papers are indirectly linked. Cascaded bipartite networks allow for the investigation of these types of networks. Given a cascade of bipartite networks with occurrence matrices $\mathbf{O}[x_1;x_2]$, $\mathbf{O}[x_2;x_3]$, …, $\mathbf{O}[x_{n-1};x_n]$, this cascade can be reduced to a single bipartite network with occurrence matrix $\mathbf{O}[x_1;x_n]$ listing the link weights between the x_1 entities and the x_n entities in the network.

Consider a pair of cascaded bipartite networks, with entity-types x_1, x_2, and x_3, as shown in Figure 5.

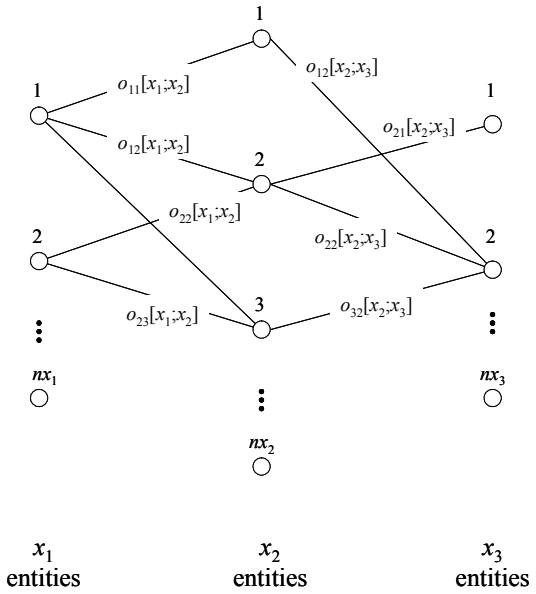


Figure 5. Diagram of a pair of cascaded bipartite networks.

There are nx_1, nx_2, and nx_3 entities of the entity-types x_1, x_2, and x_3 respectively. A pair of links that connects an x_1 entity to an x_3 entity is defined as a *path*. Figure 6, part (a) shows a path from x_1 entity i to x_3 entity j, connected through x_2 entity k by links $o_{ik}[x_1;x_2]$ and $o_{kj}[x_2;x_3]$. There are nx_2 possible paths from x_1 entity i to x_3 entity j as shown in Figure 6 part (b).

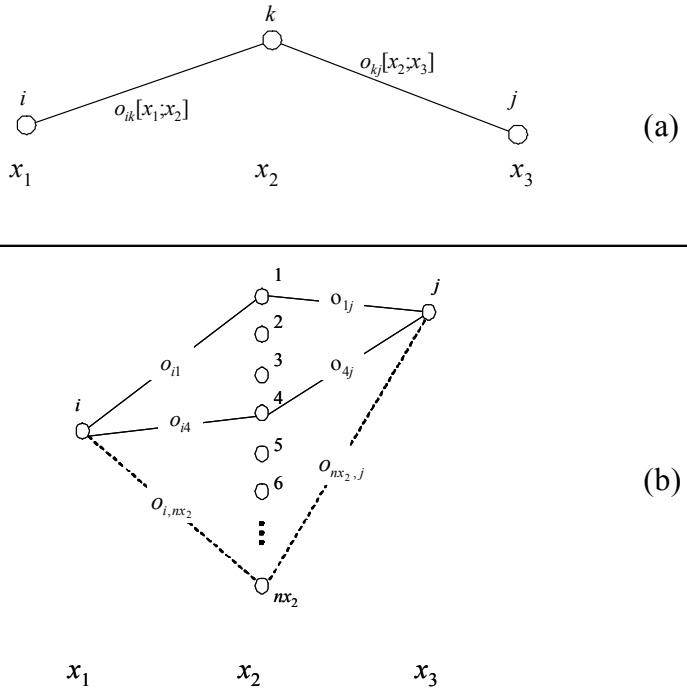


Figure 6. Paths between x_1 entity i and x_3 entity j though x_2 entities.

The *path weight* associated with a path is calculated from the weights of the path's two links using a *path weight function*:

$$p_{ij}(k) = \mathrm{f}_2(o_{ik}[x_1;x_2], o_{kj}[x_2;x_3]) \tag{2}$$

The resulting link weight from x_1 entity i to x_3 entity j is calculated from the path weights of all possible paths between those two entities using a *path combining function*:

$$o_{ij}[x_1;x_3] = \mathrm{f}_1\left[\, p_{ij}(1), p_{ij}(2), \ldots, p_{ij}(nx_2) \right] \tag{3}$$

Substituting Equation (2) into Equation (3) gives the *link weight function* which defines the rules for calculating link weights of cascaded bipartite networks:

$$o_{ij}[x_1;x_3] = \mathrm{f}_1\left[\mathrm{f}_2\left(o_{i1}, o_{1j} \right), \mathrm{f}_2\left(o_{i2}, o_{2j} \right), \ldots \mathrm{f}_2\left(o_{inx_2}, o_{nx_2 j} \right) \right] \tag{4}$$

where the first of each of the f_1 operands is taken from occurrence matrix $\mathbf{O}[x_1;x_2]$, and the second of each of the f_2 operands is taken from occurrence matrix $\mathbf{O}[x_2;x_3]$.

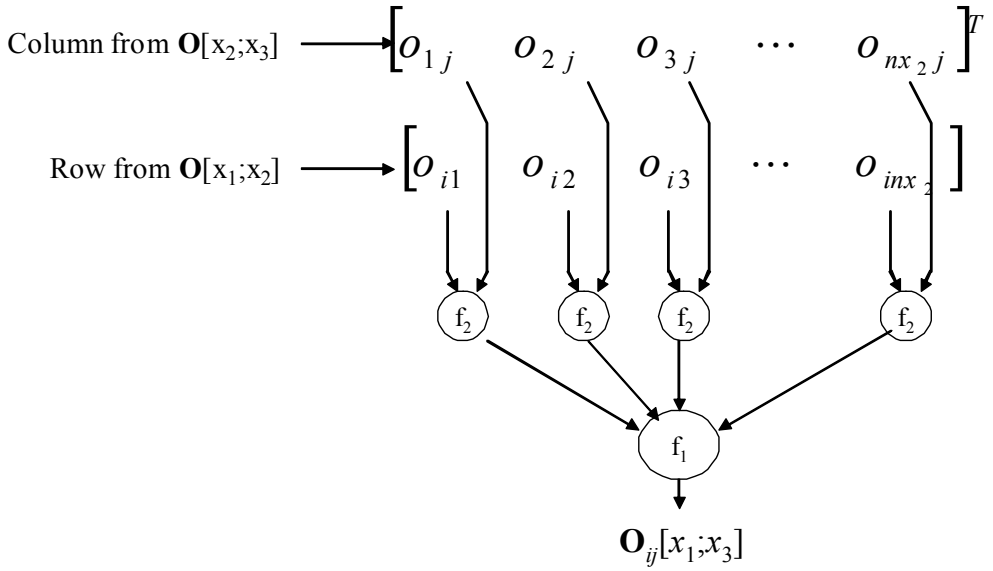


Figure 7. Diagram illustrating vector operation of the link weight function.

The link weight function of Equation (4) is a matrix function that is used to compute all the nx_1 times nx_3 possible weights of the occurrence matrix $\mathbf{O}[x_1;x_3]$ according to the rules for weight computation given by $f_1(i,j)$ and $f_2(i,j,k)$. Consider Figure 7 which illustrates how the link weight function uses row i of $\mathbf{O}[x_1;x_2]$ and column j of $\mathbf{O}[x_2;x_3]$ to produce element o_{ij} of matrix $\mathbf{O}[x_1;x_3]$. As shown, the function f_2 is applied to matching elements of the row vector and column vector to produce nx_2 scalar results. The function f_1 operates on all these nx_2 results to produce the final scalar result $o_{ij}[x_1;x_3]$.

Table 2. List of function pairs f_1 and f_2 that can be used in the link weight function for various applications. Matrix **A** describes the first bipartite network and matrix **B** describes the second bipartite network.

	f_1	f_2	**function**	**application**
1	$\sum_{k=1}^{nx_2} f_2\left(a_{ik}, b_{kj}\right)$	$a_{ik} \cdot b_{kj}$	matrix multiplication	calculate occurrence and co-occurrence counts for when **A** or **B** is binary
2	$\sum_{k=1}^{nx_2} f_2\left(a_{ik}, b_{kj}\right)$	$\left(a_{ik}^{-r} + b_{kj}^{-r}\right)^{-1/r}$	inverse Minkowski	calculate links similar to conductances in series
3	$\sum_{k=1}^{nx_2} f_2\left(a_{ik}, b_{kj}\right)$	$\min\left(a_{ik}, b_{kj}\right)$	overlap function	calculate co-occurrence counts and indirect occurrence counts
4	$\min\left[\sum_{k=1}^{nx_2} f_2\left(a_{ik}, b_{kj}\right), 1\right]$	$a_{ik} \cdot b_{kj}$	simple occurrence/co-occurrence	produce binary occurrence/co-occurrence matrix
5	$\max[f(a_{i1}, b_{1j}), \ldots f(a_{inx_2}, b_{nx_2 j})]$	$\left(a_{ik}^{-r} + b_{kj}^{-r}\right)^{-1/r}$	maximum similarity path of length 2 from node i to node j.	pathfinder analysis (see Chapter 13.11)
6	$\max[f(a_{i1}, b_{1j}), \ldots f(a_{inx_2}, b_{nx_2 j})]$	$\min\left(a_{ik}, b_{kj}\right)$	maximum similarity path of length 2 from node i to node j for $r = -\infty$	pathfinder analysis (see Chapter 13.11)

The link weight function can be used not only for calculating occurrence matrices of cascaded bipartite networks, but it is also useful for calculating co-occurrence matrices (see Chapter 5) and for performing pathfinder network calculations (See Chapter 13.11). Table 2 shows a list of various functions f_1 and f_2 that can be used for the link weight function. For applications where at least one of the matrix arguments is binary, standard matrix multiplication, the first function listed in Table 2, is often used because it directly yields simple occurrence and co-occurrence counts.

3.3 Matrix multiplication

If the path weight function f_2 is defined as a product:

$$f_2\left(o_{ik}[x_1;x_2],o_{kj}[x_2;x_3]\right)=o_{ik}[x_1;x_2]\cdot o_{kj}[x_2;x_3] \tag{5}$$

and the path combining function f_1 is a summation:

$$\begin{aligned} &f_1\left[f_2\left(o_{i1}[x_1;x_2],o_{1j}[x_2;x_3]\right),\ldots,f_2\left(o_{inx_2}[x_1;x_2],o_{nx_2j}[x_2;x_3]\right)\right] \\ &\qquad=\sum_{k=1}^{nx_2} f_2\left(o_{ik}[x_1;x_2],o_{kj}[x_2;x_3]\right). \end{aligned} \tag{6}$$

Then the link weight function is simply standard matrix multiplication:

$$o_{ij}[x_1;x_3]=\sum_{k=1}^{nx_2} o_{ik}[x_1;x_2]\cdot o_{kj}[x_2;x_3]\,. \tag{7}$$

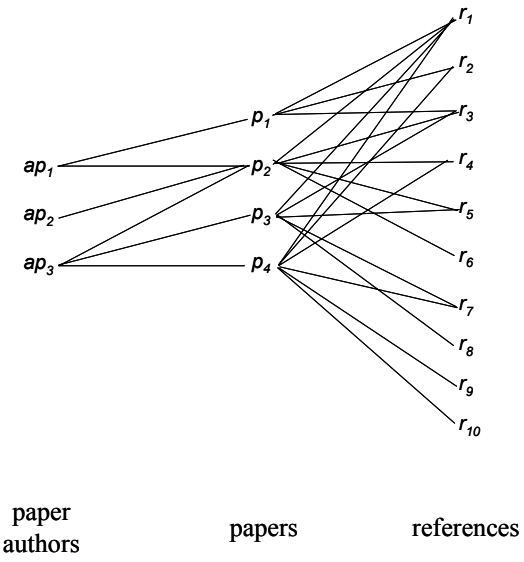


Figure 8. Example of cascaded bipartite networks: paper author to paper network cascaded with paper to reference network. All links have unity weight.

As an example, assume that x_1, x_2, and x_3 are paper authors, papers and references respectively, as shown in Figure 8. The binary matrix **O**[*ap*;*p*] lists the associations of the individual paper authors with each paper:

$$\mathbf{O}[ap;p]=\begin{bmatrix}1&1&0&0\\0&1&0&0\\0&1&1&1\end{bmatrix} \tag{8}$$

and the binary matrix **O**[*p*;*r*] lists the associations of individual papers with each reference:

$$\mathbf{O}[p;r]=\begin{bmatrix}1&1&1&0&0&0&0&0&0&0\\1&0&1&1&1&1&0&0&0&0\\1&0&1&0&1&0&1&1&0&0\\1&1&0&1&0&0&1&0&1&1\end{bmatrix}. \tag{9}$$

It is easy to show that in the case of binary matrices, using matrix multiplication as the link weight function will generate counts of associations between entities of the first and third entity-types in the cascade pair of bipartite networks: Using matrix multiplication:

$$\mathbf{O}[ap;r]=\mathbf{O}[ap;p]\cdot\mathbf{O}[p;r] \tag{10}$$

This yields:

$$\begin{aligned}\mathbf{O}[ap;r]&=\begin{bmatrix}1&1&0&0\\0&1&0&0\\0&1&1&1\end{bmatrix}\begin{bmatrix}1&1&1&0&0&0&0&0&0&0\\1&0&1&1&1&1&0&0&0&0\\1&0&1&0&1&0&1&1&0&0\\1&1&0&1&0&0&1&0&1&1\end{bmatrix}\\&=\begin{bmatrix}2&1&2&1&1&1&0&0&0&0\\1&0&1&1&1&1&0&0&0&0\\3&1&2&2&2&1&2&1&1&1\end{bmatrix}.\end{aligned} \tag{11}$$

This is a matrix, **O**[*ap*;*r*], containing the counts of occurrences of individual paper authors to individual references. When cascading bipartite networks where, for both networks, the occurrence matrix weights correspond to association counts, and where one of the occurrence matrices is binary, it can be shown that using matrix multiplication as the link weight function yields an occurrence matrix whose weights are counts of occurrences from entities of the first entity-type to the entities of the third entity-type.

3.4 Overlap function

While matrix multiplication is often used as the link weight function, in the situation where both matrices are non-binary, matrix multiplication fails to give meaningful occurrence counts when used as a link weight function. The raw weights of the adjacency matrices in collections of papers are usually associated with occurrence counts, and the occurrence counts are binary, that is, there is either no association between a pair of entities and their corresponding weight is zero, or there is a single association between the pair of entities and the weight is unity. It can be shown that calculation of link weights in a cascade of bipartite networks with binary link weights will always yield the number of occurrences between any two unlike entities in the network.

When the occurrence counts are not unity then the path weights from x_1 node i to x_3 node j through x_2 node k is a function of the two links weights, o_{ik} from node i to node k and o_{kj} from node k to node j. It is natural to think of linkages as analogous to electrical conductances between entities. In this sense when thinking of links connected in successive links in a path, the resulting link weight should be limited by the smallest link weight in the path. This can be accomplished using a path weight function that finds the minimum of the weights of the two links on the path:

$$\mathrm{f}_2 = \min\left(o_{ik}[x_1;x_2], o_{kj}[x_2;x_3] \right). \tag{12}$$

Using a path combining function that sums the path weights:

$$\mathrm{f}_1 = \sum_{k=1}^{nx_2} \mathrm{f}_2\left(o_{ik}[x_1;x_2], o_{kj}[x_2;x_3] \right). \tag{13}$$

This yields the *overlap function* (Salton, 1971) as the link weight function:

$$o_{ij}[x_1;x_3] = \sum_{k=1}^{nx_2} \min\left(o_{ik}[x_1;x_2], o_{kj}[x_2;x_3] \right). \tag{14}$$

This can be defined as a matrix operation “OVL”:

$$\mathbf{O}[x_1;x_3] = OVL(\mathbf{O}[x_1;x_2], \mathbf{O}[x_2;x_3]) \tag{15}$$

The overlap function is entry 3 in Table 2. Discussion of the application and characteristics of this function can be found in Jones and Furnas (1987).

As an example, assume that x_1, x_2, and x_3 are terms, papers and reference authors respectively, as shown in Figure 9.

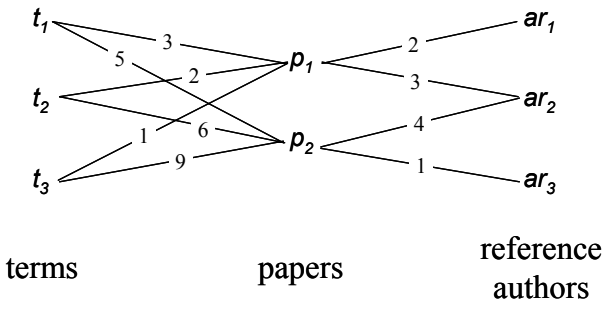


Figure 9. Example of cascaded bipartite networks with non-binary links. Terms to paper network cascaded with paper to reference author network.

The matrix $\mathbf{O}[t;p]$ lists the occurrence counts of the individual terms with each paper:

$$\mathbf{O}[t;p] = \begin{bmatrix} 3 & 5 \\ 2 & 6 \\ 1 & 9 \end{bmatrix} \tag{16}$$

and the matrix $\mathbf{O}[p;ar]$ lists the associations of individual papers with each reference author:

$$\mathbf{O}[p;ar] = \begin{bmatrix} 2 & 3 & 0 \\ 0 & 4 & 1 \end{bmatrix} \tag{17}$$

Using the overlap function to calculate the link weights of $\mathbf{O}[t;ar]$:

$$\mathbf{O}[ap;r] = \mathbf{O}[ap;p] \cdot \mathbf{O}[p;r] \tag{18}$$

This yields:

$$\mathbf{O}[t;ar] = OVL\left(\begin{bmatrix} 3 & 5 \\ 2 & 6 \\ 1 & 9 \end{bmatrix}, \begin{bmatrix} 2 & 3 & 0 \\ 0 & 4 & 1 \end{bmatrix}\right) = \begin{bmatrix} 2 & 7 & 1 \\ 2 & 6 & 1 \\ 1 & 5 & 1 \end{bmatrix} \tag{19}$$

3.5 Inverse Minkowski function

The *inverse Minkowski function*, an adaptation of the well-known Minkowski distance metric (Cios, Pedrycz, & Swiniarski, 1998), can be used when it is desired to model path weights as if the link weights were electrical conductances in series. In this case use the inverse Minkowski metric as the path weight function:

$$\mathrm{f}_2 = \left[\left(o_{ik}[x_1; x_2] \right)^{-p} + \left(o_{kj}[x_2; x_3] \right)^{-p} \right]^{-\frac{1}{p}} \tag{20}$$

Where p ranges from zero to positive infinity. Note that, in contrast to the Minkowski metric as normally expressed, the exponents in the inverse Minkowski metric are negative. This function will always generate a path weight that is less than or equal to the smallest link weight in the path, modeling a situation where indirect links tend to be weaker than direct links. Figure 10 shows a plot of the inverse Minkowski metric for the ratio of the two weights used as arguments to the function. Using a path combining function that sums the path weights:

$$\mathrm{f}_1 = \sum_{k=1}^{nx_2} \mathrm{f}_2 \left(o_{ik}[x_1; x_2], o_{kj}[x_2; x_3] \right) \tag{21}$$

Yields the final inverse Minkowski link weight function:

$$o_{ij}[x_1; x_3] = \sum_{k=1}^{nx_2} \left[\left(o_{ik}[x_1; x_2] \right)^{-p} + \left(o_{kj}[x_2; x_3] \right)^{-p} \right]^{-\frac{1}{p}} \tag{22}$$

This corresponds to entry 2 of Table 2, and can be defined as a matrix operation "INVMINK":

$$\mathbf{O}[x_1; x_3] = INVMINK(\mathbf{O}[x_1; x_2], \mathbf{O}[x_2; x_3]) \tag{23}$$

When this function is used with $p = \infty$, Equation (20) produces the minimum of its arguments and so reverts to Equation (12), making the inverse Minkowski link weight function revert to the overlap link weight function. When $p = 1$, then the path weight function, Equation (20), becomes:

$$\mathrm{f}_2 = \left[\frac{1}{o_{ik}[x_1; x_2]} + \frac{1}{o_{kj}[x_2; x_3]} \right]^{-1} \tag{24}$$

This makes the path weight function produce a value that is twice the harmonic average of the link weights of the path. This is equivalent to calculating the path weight by modeling the link weights as electrical conductances in series.

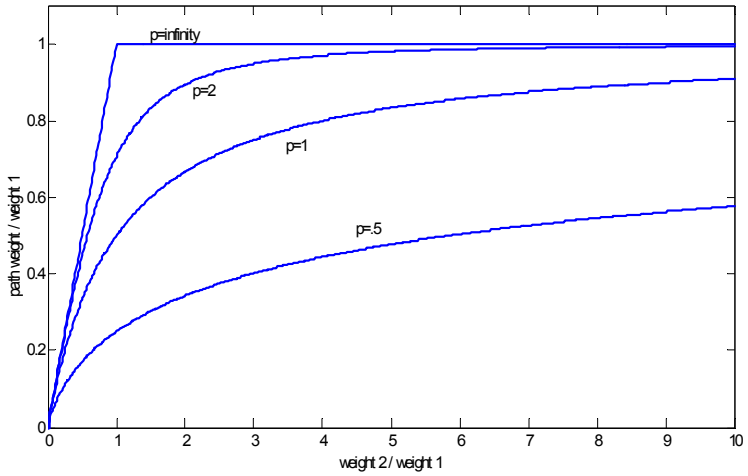


Figure 10. Plot of inverse Minkowski metric for various values of exponent *p* as a function of the ratio of weights. Note that when *p* is infinity the inverse Minkowski metric reverts to the min function. When *p* = 1 the Minkowski metric yields two times the harmonic mean of the two weights.

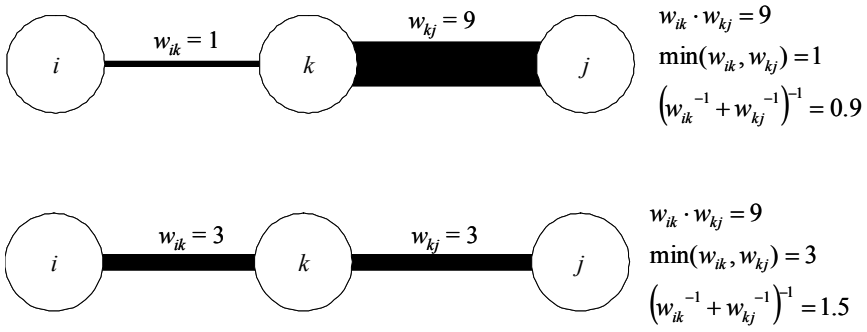


Figure 11. Illustration of the computation of link weights using different path weight functions.

Figure 11 presents a diagram showing the calculation of path weights using three different path weight functions. The upper diagram shows two entities connected through a third entity with links that are greatly unequal. The lower diagram shows entities connected by links that are of equal weight. Using multiplication as the path weight function generates path weights that are the same for both these cases even though it is logical that the upper path weight should be less than the lower path weight, since it has one link of low value. Using the minimum as the path weight function yields a smaller weight for the upper path compared to the lower path but doesn't reduce the path weights further. Using the inverse

Minkowski function as path weight function lowers each path weight below its minimum link weight, thus reducing the link weight for an indirect link over that possible using a direct connection, similar to conductances in series.

The concepts introduced in this chapter: 1) bipartite networks of entities, 2) cascaded bipartite networks, and 3) link weight functions, provide a systematic means of tracing indirect links between entities in collections of papers and calculating the weights of those indirect links. The framework provided by the ER model and dyadic notation introduced in Chapter 2 and the indirect link computation methods introduced in this chapter, are the foundation upon which the proposed mathematical treatment of collections of papers is built. This mathematical treatment is a matrix based formulation based on occurrence matrices, discussed in Chapter 4, listing the links between unlike entities.

4. OCCURRENCE MATRICES

4.1 Definition of occurrence matrix

As discussed in Chapter 3.1 an occurrence matrix lists the weights of links in a bipartite network. This chapter discusses occurrence matrices in a less general sense, as integer matrices that count associations between dyads of unlike entities. Define an occurrence matrix as $\mathbf{O}[x_1;x_2]$, where $o_{ij}[x_1;x_2]$ is equal to the number of times that primary entity i is associated with relative entity j. The occurrence matrix $\mathbf{O}[x_1;x_2]$ is an nx_1 by nx_2 matrix whose rows correspond to the primary entities, of entity-type x_1, and whose columns correspond to the relative entities, of entity-type x_2. As an example, consider a collection of 4 papers and 10 references and its corresponding paper to reference matrix:

$$\mathbf{O}[p;r] = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \tag{25}$$

In this case $\mathbf{O}[p;r]$ is a matrix whose rows correspond to papers and whose columns correspond to references. Occurrence matrices for direct relations are usually binary because associations among unlike entities in the collection are usually counted as logical associations. Paper author to paper associations, for example, are logical, i.e., given an author and a paper, the author is either an author of the paper or not an author of the paper. This is coded in the occurrence matrix as 1 or 0 for associated or not associated respectively. In another example, while it is true that a paper can cite a reference multiple times, generating multiple associations for paper-reference dyads, abstract services and citation index services never record the number of times that references are cited in papers, and report only the logical association of "cited or not cited," resulting in a binary paper to reference matrix. Associations of linguistic terms with papers however, are often extracted as term counts from titles, abstracts, and paper text bodies, and so produce non-binary occurrence matrices. Excepting the paper to term relationship, all of the 7 direct relationships shown in the entity relationship diagram of Figure 1 are assumed to be binary occurrence matrices.

Occurrence matrices for indirect relations are often not binary matrices. For example in paper-reference author dyads, a reference author may appear in multiple references associated with a paper, resulting in multiple associations of reference authors with papers, giving a non-binary occurrence matrix. Using a threshold value, non-binary occurrence matrices can be converted to binary matrices.

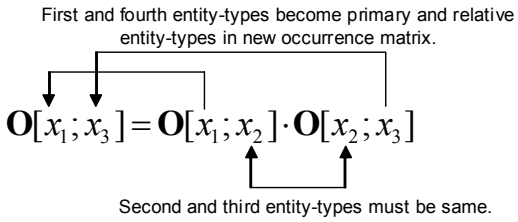
Note the following property of occurrence matrices:

$$\mathbf{O}[x_2; x_1] = \mathbf{O}[x_1; x_2]^T \tag{26}$$

Using dyad identifier notation, transposing the variables is equivalent to transposing the occurrence matrix.

4.2 Indirect occurrence matrices calculated by cascading occurrence matrices

As discussed in Chapter 3.3, in a cascade of bipartite networks where all the link weights are binary, the use of matrix multiplication as the link weight function results in computing an occurrence matrix where the link weights correspond to the counts of associations between starting and ending entities in the cascade. Using the dyad identifier and applying Equation 26 when appropriate, the matrix multiplication is expressed as:



(27)

Note that, with the notation as defined, the primary entity-type of the resulting indirect occurrence matrix is the primary entity-type of the first matrix on the right side, the relative entity-type of the resulting occurrence matrix is the relative entity-type of the second matrix on the right side. The first matrix relative entity-type and the second matrix primary entity-type must be the same. That is, the inner two entity-types in the four entity-types that appear on the right side are the same. Thus, the dyad identifier notation makes it quite easy to arrange occurrence matrices for calculation of indirect occurrence matrices. In fact, several occurrence matrices may be multiplied together at once to calculate an indirect occurrence matrix for several cascaded bipartite networks. The matrices are arranged so that desired primary entity-type is the primary entity-type of the first matrix, the desired relative entity-type is the relative entity-type of the last matrix, and all the primary entity-types of the matrices within this order are matched by the same relative entity-type in the matrix proceeding it. Assume the collection of 4 papers containing 10 references from Equation (25) has 3 paper authors, and that the paper to paper author matrix is given by:

$$\mathbf{O}[p; ap] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \tag{28}$$

Also assume that the collection of papers has 6 reference authors and that the reference to reference author matrix is given by:

$$\mathbf{O}[r;ar] = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{29}$$

Now suppose we wish to find the paper author to reference author matrix. Consulting Figure 1, the direct links from paper authors to reference authors go from paper author to paper to reference to reference author. Calculation of the occurrence matrix, **O**[*ap*;*ar*], from paper author to reference author is performed by the matrix multiplication:

$$\mathbf{O}[ap;ar] = \mathbf{O}[ap;p] \cdot \mathbf{O}[p;r] \cdot \mathbf{O}[r;ar]$$

desired primary entity-type

relative entity-type of preceeding matrix matched to primary entity-type of following matrix

desired relative entity-type

(30)

First find the paper author to reference matrix by pre-multiplying the paper author to paper matrix by the paper to reference matrix:

$$\mathbf{O}[ap;r] = \mathbf{O}[ap;p] \cdot \mathbf{O}[p;r] =$$

$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 2 & 2 & 2 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 2 & 1 & 1 & 1 & 0 & 1 & 1 \end{bmatrix} \tag{31}$$

Then pre-multiply the paper author to reference matrix against the reference to reference author matrix:

$$\mathbf{O}[ap;ar] = \mathbf{O}[ap;r] \cdot \mathbf{O}[r;ar] =$$

$$\begin{bmatrix} 3 & 2 & 2 & 2 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 2 & 1 & 1 & 1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 4 & 3 & 2 & 1 & 1 \\ 1 & 2 & 0 & 0 & 0 & 0 \\ 2 & 2 & 3 & 2 & 1 & 1 \end{bmatrix} \tag{32}$$

The result in Equation 32 gives the desired occurrence matrix of paper authors to reference authors for the example. Matrix multiplication provides an easy method of calculating occurrence matrices for indirect relations. This is especially useful when using existing sparse matrix multiplication algorithms for making the computations.

4.3 Equivalence matrices

In some types of bibliometric studies, particularly studies of information flow from paper to paper through citations (Garfield et al., 2003), it is useful to find and list those pairs of bibliometric entities that correspond to the same physical entity. This can be done by matching attributes of the entities, for example, matching references and papers that have the same first author, publication year, volume, journal, issue and page number. Paper journals and reference journals, and additionally, paper authors and reference authors can be matched by matching names. Note that in a paper collection there are typically 10 times more references, reference authors, and reference journals than there are papers, paper authors, and paper journals respectively. A paper collection will typically have many references that correspond to books, and these references will have no corresponding physical paper. So the list of references and papers that correspond to the same physical paper is typically very incomplete. Many references will have no corresponding papers and many papers will have no corresponding references. Define an equivalence matrix that maps the pairs of entities that correspond to the same physical entity:

$$\mathbf{A}[cx_1, x_2] : a_{ij}[cx_1; x_2] = \begin{cases} 1, & \text{if } x_1 \text{ corresponds to } x_2 \\ 0, & \text{otherwise} \end{cases} \tag{33}$$

Here the variable cx_1 is used to denote "x_1 as a *cited* entity." As an example, assume that in the collection of papers denoted by Equation (25) that three of the papers can be matched to references, and that this yields the following equivalence matrix:

$$\mathbf{A}[cp, r] = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{34}$$

The adjacency matrix that relates direct citation relations among like entities can be calculated using the matrix multiplication:

$$\mathbf{O}[x_1, cx_1] = \mathbf{O}[x_1, x_2] \cdot \mathbf{A}[x_2, cx_1] \tag{35}$$

Using the example of four papers from Equation (25) and its equivalence matrix from Equation (34), the adjacency matrix from papers to cited papers is computed:

$$\mathbf{O}[p; cp] = \mathbf{O}[p; r] \cdot \mathbf{A}[r; cp] =$$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix} \tag{36}$$

Figure 12 shows a diagram of the calculation of an adjacency matrix of papers to papers using an equivalence matrix.

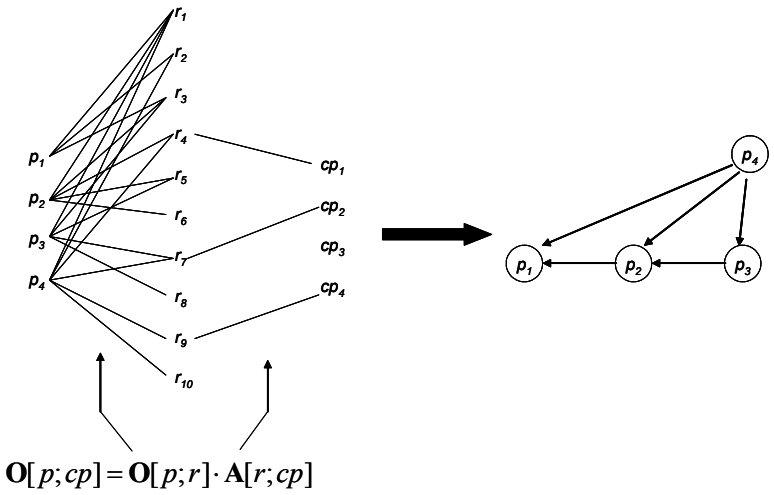


Figure 12. Diagram showing calculation of paper to paper adjacency matrix using an equivalence matrix.

4.4 Membership matrices

Given a membership function for an entity-type, a membership matrix lists the association of the entities with each of the groups of entities. We can use the letter 'G' (for 'group') to denote the membership matrix: $\mathbf{G}[gx_1, x_1]$. For example, assume that papers from the example of Equation (25) are clustered into two research fronts and that the membership matrix for the research front is given by:

$$\mathbf{G}[gp, p] = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \tag{37}$$

which shows that papers 1 and 2 are members of paper group 1 and that papers 3 and 4 are members of paper group 2. Also assume that references were clustered into 3 reference groups, and that the membership matrix for reference groups is:

$$\mathbf{G}[gr, r] = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{38}$$

Matrix multiplication can be used to show the association of entities of other entity-types with the groups:

$$\mathbf{O}[gx_1, x_2] = \mathbf{G}[gx_1, x_1] \cdot \mathbf{O}[x_1, x_2] \tag{39}$$

Using our example we can compute the relationship of paper groups to base references through matrix multiplication:

$$\mathbf{O}[gp;gr] = \mathbf{G}[gp;p]\cdot\mathbf{O}[p;r]\cdot\mathbf{G}[r;gr] =$$

$$\begin{bmatrix}1&1&0&0\\0&0&1&1\end{bmatrix}\begin{bmatrix}1&1&1&0&0&0&0&0&0&0\\1&0&1&1&1&1&0&0&0&0\\1&0&1&0&1&0&1&1&0&0\\1&1&0&1&0&0&1&0&1&1\end{bmatrix}\begin{bmatrix}1&0&0\\1&0&0\\1&0&0\\0&1&0\\0&0&1\\0&1&0\\0&0&1\\0&1&0\\1&0&0\\0&1&0\end{bmatrix} = \begin{bmatrix}5&2&1\\5&3&3\end{bmatrix} \tag{40}$$

Membership matrices can also be used to clean up matrices after disambiguating names and references. Suppose that in the example of Equation (25) after disambiguating references it's found that reference 5 and reference 2 refer to the same physical entity. Then, in order to eliminate reference 5 and move its citations over to reference 2, construct a membership matrix:

$$\mathbf{G}[r1;r] = \begin{bmatrix}1&0&0&0&0&0&0&0&0&0\\0&1&0&0&1&0&0&0&0&0\\0&0&1&0&0&0&0&0&0&0\\0&0&0&1&0&0&0&0&0&0\\0&0&0&0&0&1&0&0&0&0\\0&0&0&0&0&0&1&0&0&0\\0&0&0&0&0&0&0&1&0&0\\0&0&0&0&0&0&0&0&1&0\\0&0&0&0&0&0&0&0&0&1\end{bmatrix} \tag{41}$$

where $r1$ denotes the collection after reference 5 has been eliminated. The new paper reference can be found from:

$$\mathbf{O}[p;r1] = \mathbf{O}[p;r]\cdot\mathbf{G}[r;r1] \tag{42}$$

Which gives:

$$\mathbf{O}[p;r1] = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \tag{43}$$

Occurrence matrices are the building blocks of the mathematical treatment proposed here. They are lists of the links between entities in the collection of papers and so form the basic tools for analysis of structure of the complex networks within paper collections that are manifestations of social process and scientific progress of research in a research specialty.

In review of topics discussed so far in this report: 1) from Chapter 2, the entity-relationship (ER) model of paper collections allows identification of objects (entities) to be studied in the collections, and shows how those entities are directly linked. The ER model further shows the chains of links that must be followed when computing indirect links. 2) from Chapter 3, the cascaded bipartite network model of indirect links provides a general method of tracing and computing indirect links within the collection of papers. This method is general and applicable to both binary and non-binary links and adaptable to most methods currently in use for calculating link weights. 3) occurrence matrices list all the links between entities of interest and allow simple manipulation of the those links using matrix arithmetic.

The next chapter will introduce co-occurrence matrices, which are used to measure links between like entities. It will be shown that co-occurrence matrices are easily found from occurrence matrices using simple matrix arithmetic and cascaded bipartite network models. Co-occurrence matrices form the basis of clustering of like entities, which is so important for mapping the social and knowledge structures that are manifested in collections of journal papers.

5. CO-OCCURRENCE MATRICES

5.1 Definition of co-occurrence matrix

Co-occurrence matrices are symmetric matrices that list the number of the co-occurrence counts of common associations that pairs of like primary entities have with entities of some relative entity-type. For example, the *co-occurrence matrix of papers relative to references* lists the number of common references for each pair of papers in the collection of papers. For binary occurrence matrices the co-occurrence matrix can be found by post multiplying the occurrence matrix by its transpose:

$$\mathbf{C}[x_1;x_2] = \mathbf{O}[x_1;x_2]\cdot\mathbf{O}[x_2;x_1] \tag{44}$$

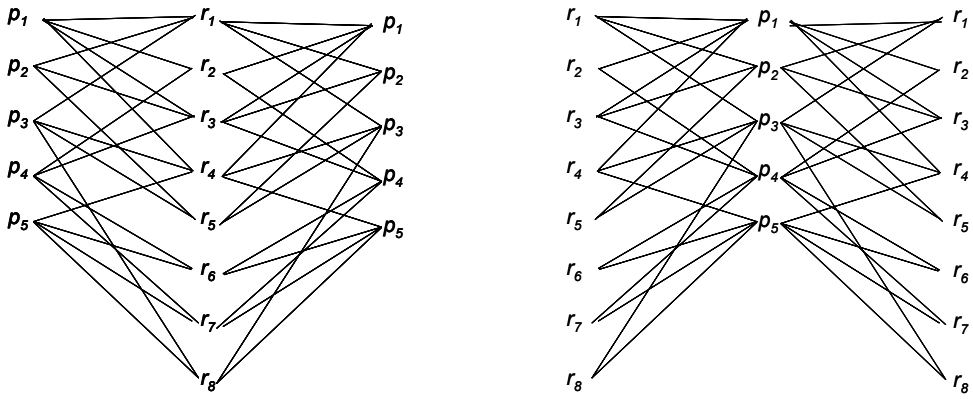
Where $\mathbf{C}[x_1;x_2]$ is the co-occurrence matrix listing the number of common associations of pairs of x_1 entities with x_2 entities. For example, to calculate the co-occurrence of papers relative to papers (bibliographic coupling) using the paper to reference matrix example from Equation (25):

$$\mathbf{C}[p;r] = \mathbf{O}[p;r]\cdot\mathbf{O}[r;p] =$$

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 2 & 2 & 2 \\ 2 & 5 & 3 & 2 \\ 2 & 3 & 5 & 2 \\ 2 & 2 & 2 & 6 \end{bmatrix} \tag{45}$$

The diagonal of the co-occurrence matrix $c_{ii}[x_1; x_2]$ lists the number of associations that each x_1 has with entities of the x_2 entity-type. For example in the bibliographic coupling matrix, $\mathbf{C}[p;r]$, calculated in Equation (45), the diagonal lists the number of references each papers cites.

Computation of co-occurrences can be viewed, similar to the discussion of Chapter 3.2, as the calculation of links in a cascade of two bipartite networks. Given a bipartite network of two unlike entity-types, mirror the network across the relative entity-type entities to obtain a cascade of two networks. For example, the paper to reference network shown in Figure 3 has been mirrored on the references to produce the paper-reference-paper cascade of two bipartite networks shown in Figure 13 (a). Calculating the path weights of this 'virtual' cascade using matrix multiplication will produce the co-occurrence counts of papers relative to references (bibliographic coupling) as was done in Equation (45).



(a) Bibliographic coupling (b) Co-citation

Figure 13. Mirror of paper to reference bipartite network to calculate co-occurrence as a cascade of two bipartite networks. (a) Mirror across references to calculate bibliographic coupling. (b) Mirror across papers to calculate co-citation.

The same network of Figure 3 can be mirrored on the papers to produce the reference-paper-reference cascade of bipartite networks shown in Figure 13 (b). Calculating the link weights in this network using matrix multiplication yields the co-occurrence counts of references relative to papers (co-citation.) Note that each occurrence matrix has two co-occurrence matrices associated with it. Figure 14 illustrates this for a sample paper to reference occurrence matrix, $\mathbf{O}[p;r]$. Note to the right of $\mathbf{O}[p;r]$ is the square symmetric bibliographic coupling matrix $\mathbf{C}[p;r]$, whose size is number of papers in $\mathbf{O}[p;r]$. Similarly, below $\mathbf{O}[p;r]$ is the square symmetric co-citation matrix, $\mathbf{C}[r;p]$ whose size is the number of references in $\mathbf{O}[p;r]$.

5.2 Overlap function for calculating co-occurrence

Linguistic term occurrence matrices are not binary since each term usually occurs multiple times in a paper. Because of this, it is not desirable to calculate term co-occurrence matrices using matrix multiplication, because the resulting link weights cannot be interpreted. Noting that calculation of co-occurrence matrices is analogous to computing link weights for a pair of cascaded bipartite networks, as was demonstrated in Figure 13 and the discussion above, link weight functions can be used to find term co-occurrence matrices. This can be done, for example, using the overlap function of Chapter 3.4.

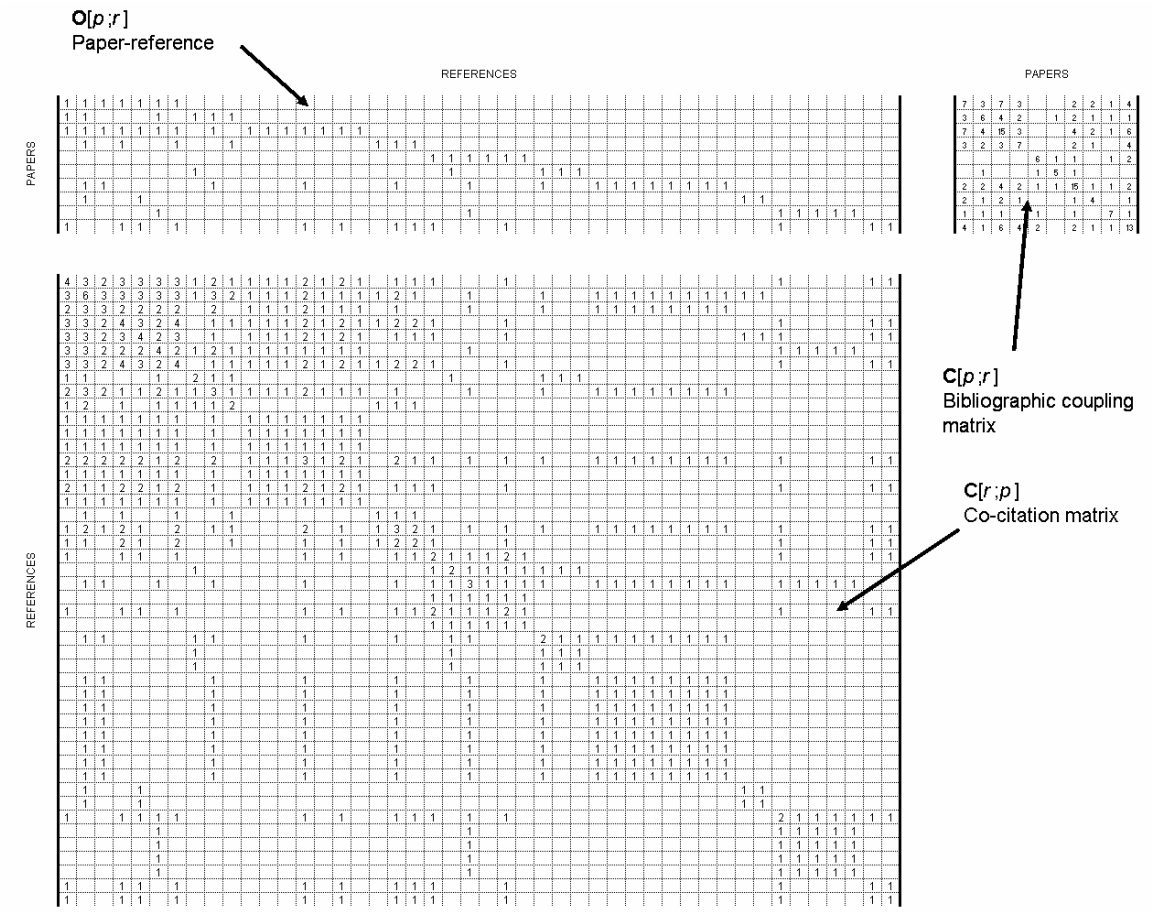


Figure 14. Diagram showing that each occurrence matrix is associated with a pair of co-occurrence matrices. Upper left matrix is paper to reference occurrence matrix **O**[*p*;*r*], below is reference co-occurrence matrix relative to papers (co-citation matrix), **C**[*r*;*p*]. Upper right matrix is paper co-occurrence matrix relative to references (bibliographic coupling matrix), **C**[*p*;*r*].

As an example, assume the paper to term matrix:

$$\mathbf{O}[p;t1] = \begin{bmatrix} 8 & 9 & 5 & 3 & 1 & 0 \\ 5 & 4 & 9 & 2 & 0 & 1 \\ 0 & 0 & 2 & 6 & 5 & 4 \\ 1 & 1 & 0 & 5 & 2 & 5 \end{bmatrix} \tag{46}$$

Using the overlap function, the term co-occurrence matrix relative to papers is found:

$$\mathbf{C}[p;t] = OVL(\mathbf{O}[p;t], \mathbf{O}[t;p]) =$$

$$OVL\left(\begin{bmatrix} 8 & 9 & 5 & 3 & 1 & 0 \\ 5 & 4 & 9 & 2 & 0 & 1 \\ 0 & 0 & 2 & 6 & 5 & 4 \\ 1 & 1 & 0 & 5 & 2 & 5 \end{bmatrix}, \begin{bmatrix} 8 & 5 & 0 & 1 \\ 9 & 4 & 0 & 1 \\ 5 & 9 & 2 & 0 \\ 3 & 2 & 6 & 5 \\ 1 & 0 & 5 & 2 \\ 0 & 1 & 4 & 5 \end{bmatrix}\right) = \begin{bmatrix} 26 & 16 & 6 & 6 \\ 16 & 21 & 5 & 5 \\ 6 & 5 & 17 & 14 \\ 6 & 5 & 11 & 14 \end{bmatrix} \quad (47)$$

5.3 Commonly used co-occurrence matrices

Bibliographic coupling matrix. Bibliographic coupling between a pair of papers was defined by Kessler (1963) as the number of references that both papers cite. Bibliographic coupling is used to cluster papers into research fronts, that is, groups of papers that cover the same topic. The bibliographic coupling matrix is a symmetric matrix that contains the bibliographic coupling counts between all papers in a collection:

$$\mathbf{C}[p;r] = \mathbf{O}[p;r] \cdot \mathbf{O}[r;p] \quad (48)$$

Co-citation matrix. Co-citation between a pair of references is defined by Small (1973) as the number of papers that cite both references. Co-citation is used to cluster references into base reference groups. The co-citation matrix is a symmetric matrix that contains all the co-citation counts among all the pairs of references in the collection of papers:

$$\mathbf{C}[p;r] = \mathbf{O}[r,p] \cdot \mathbf{O}[p,r] \quad (49)$$

Author co-citation matrix. Author co-citation between a pair of reference authors is defined by White and Griffith (1981) as the number of papers that cite both reference authors. It is used to cluster reference authors into base reference author groups. The author co-citation matrix is a symmetric matrix that contains the author co-citation counts for all pairs of reference authors in a collection of papers. To calculate the author co-citation matrix **C**[*ar*;*p*] it is first necessary to find the paper to paper author occurrence matrix from the product of the paper to reference matrix and the reference to reference author matrix:

$$\mathbf{O}[p;ar] = \mathbf{O}[p;r] \cdot \mathbf{O}[r;ra] \quad (50)$$

The paper to reference author matrix, **O**[*p;*,*ar*], is non-binary, so matrix multiplication cannot be used to find the co-occurrence matrix, **C**[*ar*;*p*]. The co-occurrence matrix can be calculated using the overlap function as defined in Chapter 3.4:

$$\mathbf{C}[ar; p] = OVL(\mathbf{O}[ar; p], \mathbf{O}[p; ar]) \tag{51}$$

Alternatively, the single co-occurrence definition of the author co-citation count between a pair of papers is the number of papers that cite both authors at least once. This can be calculated by first setting all non-zero entries in **O**[*p*;*ar*] to unity, then pre-multiplying **O**[*p*;*ar*] by its transpose. In equation form this is accomplished by:

$$\mathbf{C}[ar; p] = \max(\mathbf{O}[ar; p], 1) \cdot \max(\mathbf{O}[p; ar], 1) \tag{52}$$

where max(**X**,*m*) is defined as a function between a matrix **X**, and a scalar *m*, where every element of **X** is set to *m* if it is greater than *m*.

Journal co-citation matrix. Journal co-citation was proposed by McCain (1991). The calculation for journal citation is very similar to the calculation for author co-citation. Calculating the non-binary paper to reference journal occurrence matrix:

$$\mathbf{O}[p; jr] = \mathbf{O}[p; r] \cdot \mathbf{O}[r; jr] \tag{53}$$

the co-occurrence matrix can be calculated from the overlap function:

$$\mathbf{C}[jr; p] = OVL(\mathbf{O}[jr; p], \mathbf{O}[p; jr]) \tag{54}$$

Alternatively, the single occurrence journal co-citation matrix can be calculated:

$$\mathbf{C}[jr; p] = \max(\mathbf{O}[jr; p], 1) \cdot \max(\mathbf{O}[p; jr], 1) \tag{55}$$

Co-term matrices. The use of co-occurrence of terms in papers for bibliometrics analysis is discussed by Callon, Courtial and Laville (1991) . Co-term matrices based on co-occurrence of title or abstract linguistic terms in papers are often based on the overlap function:

$$\mathbf{C}[t; p] = OVL(\mathbf{O}[t; p], \mathbf{O}[p; t]) \tag{56}$$

Figure 15 shows an entity-relationship diagram showing useful co-occurrence relations in a collection of papers. In this diagram co-occurrence relation labels are placed next to their primary entity-type and are adjacent to a line that connects the primary entity-type to the relative entity-type. As drawn, with 7 entity-types, there are 42 possible co-occurrence relations, although, as previously mentioned, many of these relations are trivial or otherwise not useful. In the diagram, 5 co-occurrence relations, previously studied by bibliometricians, are given names commonly used in the bibliometrics literature. Other co-occurrence relations on the diagram have no commonly used label. In these cases the labeling convention "x_1 coupling by x_2" is used, where x_1 is the primary entity and x_2 is the relative entity.

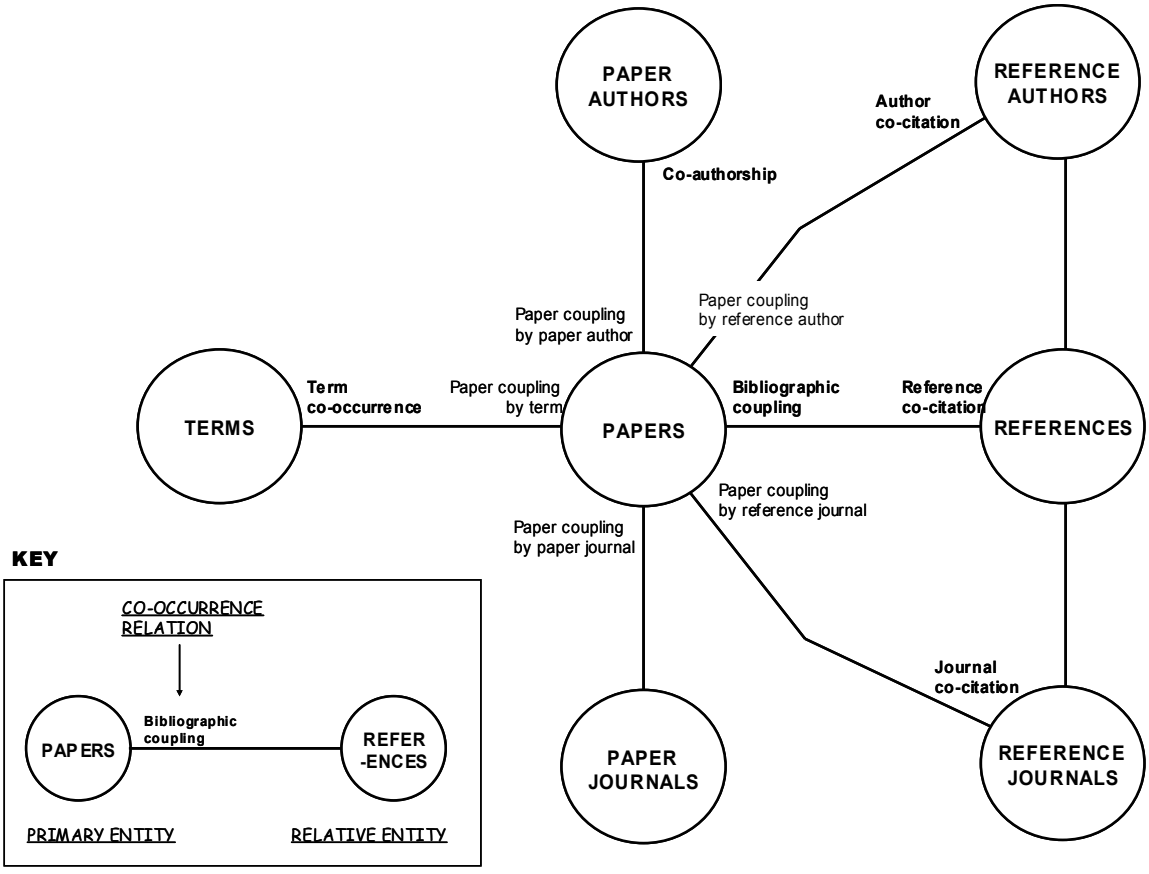


Figure 15. Entity-relationship diagram showing some useful co-occurrence relations. As shown in the diagram's key, co-occurrence relation labels are placed next to the primary entity-type and adjacent to a line connecting the primary entity-type to the relative entity-type. Co-occurrence relations shown in bold are often used by researchers in bibliometrics.

6. BIBLIOMETRIC DISTRIBUTIONS

Bibliometric distributions are useful as indicators of the processes that drive the growth of scientific research specialties and the manifestation of that growth in collections of journal papers. It is important to examine well-studied bibliometric distributions such as Lotka's Law, Bradford's, Law, and Zipf's Law in the context of the mathematical treatment proposed here. Beyond this consideration however, it is important to enumerate and classify other possibly important distributions that can occur in collections of journal papers, and discuss how those distributions can be used as indicators of important aspects of research activity in a scientific specialty.

6.1 Dyadic distributions

Dyadic distributions are distributions of the number of occurrences of entities of one entity-type relative to a second entity-type. Assume that the vector $\mathbf{M}[x_1;x_2]$ is a 1 by nx_2 vector where each element $m_i[x_1;x_2]$ is a count of the number of times x_2 entity i is associated with an x_1 entity. The vector $\mathbf{M}[x_1;x_2]$ can be considered the x_1 group to x_2 occurrence matrix that results from considering all x_1 entities as belonging to a single group. In this case, the membership matrix is a 1 by nx_1 matrix whose elements are all ones and the vector $\mathbf{M}[x_1;x_2]$ is computed as:

$$\mathbf{M}[x_1, x_2] = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \cdot \mathbf{O}[x_1, x_2] \tag{57}$$

More trivially, $\mathbf{M}[x_1;x_2]$ is the row vector of the sum of individual columns of the occurrence matrix $\mathbf{O}[x_1;x_2]$. The distribution of the elements of $\mathbf{M}[x_1;x_2]$ is the "x_1 per x_2 distribution." For example, assume that $\mathbf{M}[p;ap]$ is a row vector that lists the number of papers associated with each paper author in the collection of papers. Then the distribution of the elements of $\mathbf{M}[p;ap]$ is the "paper per paper author" distribution, the distribution that is associated with Lotka's Law.

There are two dyadic distributions for every pair of entity-types in the collection of papers. Assuming NE entity-types, there are $NE \cdot (NE\text{-}1)$ possible dyadic distributions in the collection. For the 8 entities shown in Figure 1, there are 56 possible dyadic distributions. Of these, 4 distributions have been studied extensively: 1) the paper per paper author distribution, known as Lotka's Law (Lotka, 1926), 2) the paper per paper journal distribution, known as Bradford's Law (Bradford, 1934), 3) the linguistic terms per paper distribution, known Zipf's Law (Zipf, 1949), and 4) the paper per reference distribution, referred to here as the *reference power law* (Naranan, 1971; Redner, 1998; Seglen, 1992). Of the many possible dyadic

distributions in a collection of papers, many are trivial, for example, the paper journal per paper distribution and reference journal per reference distributions are both unity, that is, there is always only one paper journal per paper, and always only one reference journal per reference. Define the dyadic x_1 per x_2 distribution as:

$$p(k, \mathbf{M}[x_1; x_2]) \tag{58}$$

This is the probability that an x_2 entity will have k associations with x_1 entities. This probability mass function can be estimated from the frequency function:

$$f(k, \mathbf{M}[x_1; x_2]) \tag{59}$$

which is the frequency table of the number of x_2 entities that have k associations with x_1 entities. This is a frequency table of the elements of the vector $\mathbf{M}[x_1;x_2]$. Figure 16 shows a diagram of dyadic distributions extracted from a paper to reference matrix, $\mathbf{O}[p;r]$. Summing the number of references along the rows of the matrix yields a column vector, the transpose of $\mathbf{M}[r;p]$, which contains the number of references cited in each paper. This can be binned into a frequency table to yield the reference per paper distribution $f(k,\mathbf{M}[r;p])$. Summing the number of papers along the columns of the matrix yields $\mathbf{M}[p;r]$, a row vector containing the number of papers citing each reference. This can be binned into a frequency table to yield the paper per reference distribution.

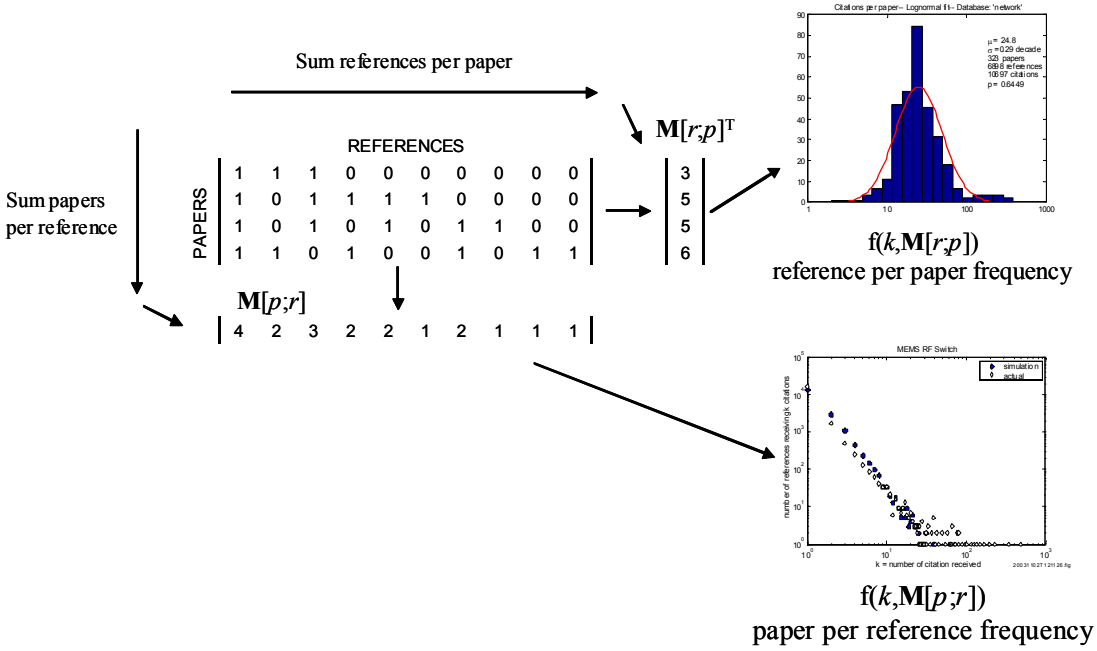


Figure 16. Diagram showing extraction of two dyadic distributions from an occurrence matrix. Using the paper to reference occurrence matrix, sum along the rows to find references per paper, then bin results to find the reference per paper distribution. Sum down the columns to find papers per reference, then bin results to get the paper per reference distribution.

Figure 17 shows an entity-relationship diagram showing several interesting dyadic distributions. Those distributions that have been well studied by bibliometricians are noted in this diagram. The typical model of the measured distribution of each dyadic distribution is noted if it has been studied in the literature or otherwise measured.

6.2 Fixed occurrence dyadic distributions

Two types of dyadic distributions are possible: 1) *fixed occurrence* distributions, and 2) *cumulating occurrence* distributions. Fixed occurrence distributions are distributions where entities of the relative entity-type can gain no additional associations as the collection grows. An example of a fixed occurrence distribution is the reference per paper distribution, $p(k,\mathbf{M}[r;p])$. All distributions with papers as the relative entity-type are fixed occurrence distributions. Individual papers are fixed and do not acquire more associations to paper authors or references as the paper collection grows. Because of this the paper author per paper and the reference per paper distributions are both fixed occurrence distributions. Fixed occurrence distributions cannot be linked to cumulative advantage (success-breeds-success) processes, since associations to the relative entities don't cumulate as the collection grows. These distributions are not as likely to be power-law distributions as the distributions associated with cumulative advantage processes. Figure 18 shows examples of two typical fixed occurrence distributions found in paper collections, a) the paper author per paper distribution, $p([k,\mathbf{M}[ap;p])$, which appears to be the result of a Poisson process, and b) the reference per paper distribution, $p([k,\mathbf{M}[r;p])$, which appears to be a log-normal distribution, typical of multiplicative noise processes.

6.3 Cumulative occurrence dyadic distributions

Cumulating occurrence distributions are associated with distributions where the relative entity-type can acquire more associations as the collection grows. All distributions where the relative entity-type is a cited entity-type are cumulating occurrence distributions. Distributions where the relative entity-type is linguistic terms are also cumulative occurrence distributions. Cumulative occurrence distributions are typically highly skewed power laws that characteristically occur for cumulative advantage processes. Figure 19 shows examples of two typical cumulating occurrence distributions found in paper collections, a) the paper per reference distribution, $p([k,\mathbf{M}[p;r])$, which is reported to be a power law (Naranan, 1971), and b) the paper per paper author distribution, $p([k,\mathbf{M}[p;ap])$, which is a power law usually referred to as Lotka's Law (1926).

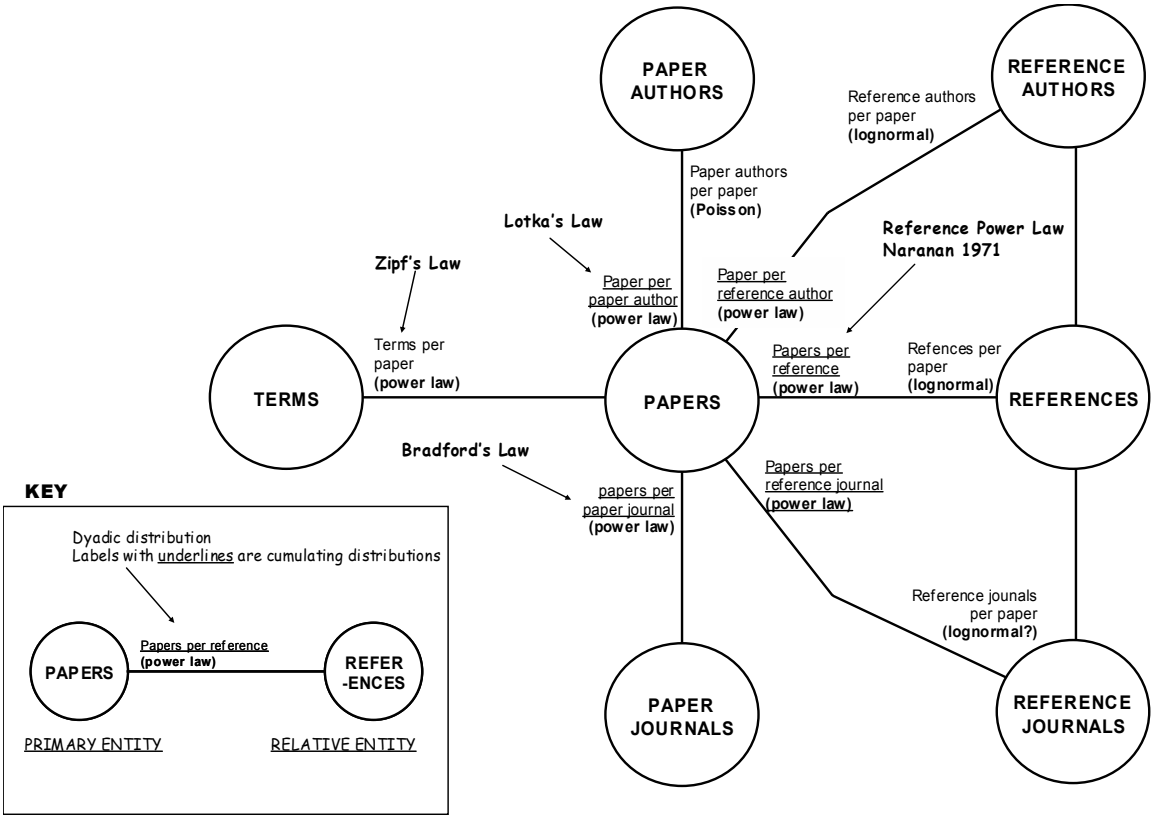
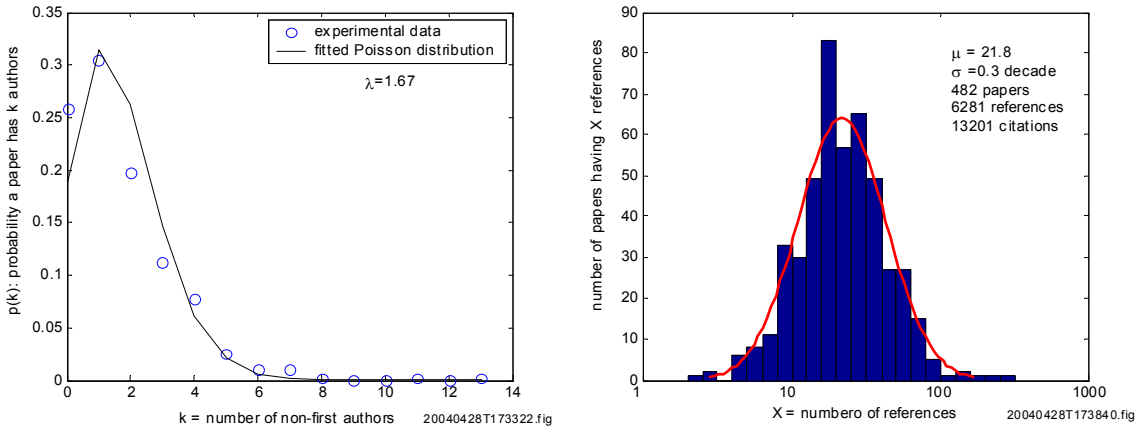


Figure 17. Entity-relationship diagram showing several interesting dyadic distributions. Distribution labels are adjacent to the distribution's primary entity-type, and adjacent to a line connecting the primary entity-type to the relative entity-type. Labels that are not underlined are fixed distributions, while labels that are underlined are cumulating distributions. Well-studied distributions are indicated by their common names in bold font.



(a) (b)

Figure 18. Example of fixed occurrence distributions. Left, (a), shows distribution of non-first authors per paper. Right, (b), shows the distribution of references per paper.

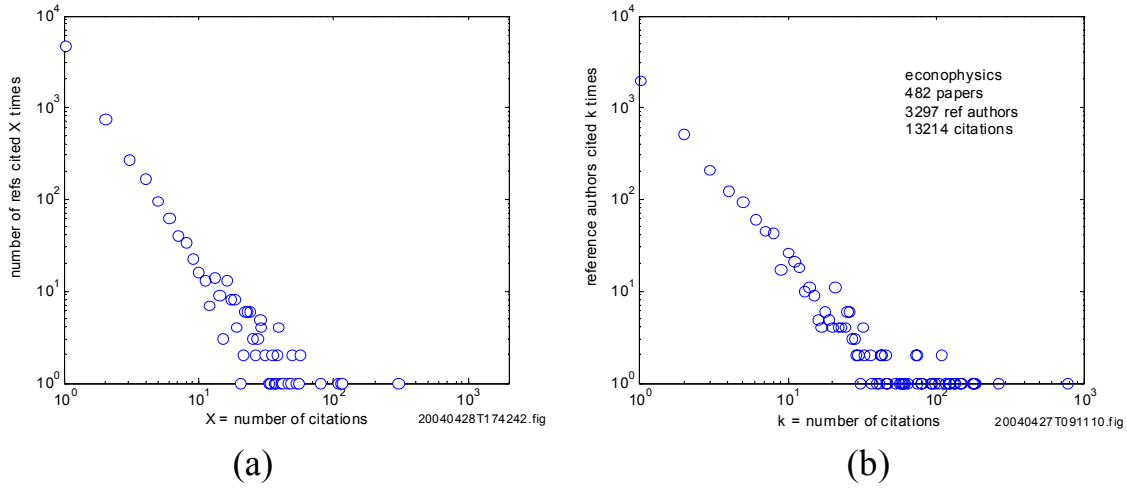


Figure 19. Example of cumulating occurrence distributions. Left, (a), shows distribution of papers per reference. Right, (b), shows distributions of citations per reference author.

6.4 Co-occurrence distributions

Co-occurrence distributions describe the probability of the number of times that a pair of entities of the primary entity-type will co-occur in their association with entities of the relative entity-type. Define the *co-occurrence of x_2 per x_1 pair distribution* as:

$$p(k, \mathbf{C}[x_1; x_2]) \tag{60}$$

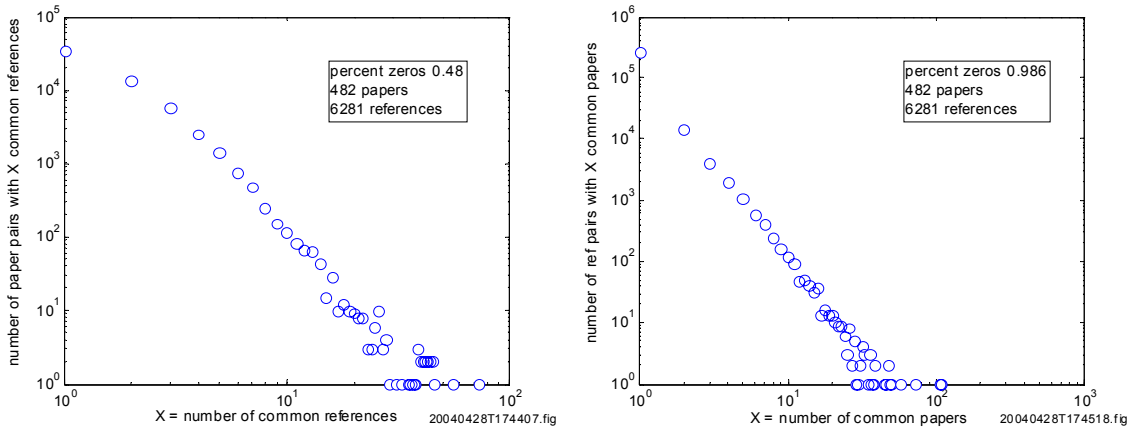
This is the probability that a pair of x_1 entities will have k co-occurrences in their associations with x_2 entities. The frequency:

$$f(k, C[x_1; x_2]) \tag{61}$$

is the frequency of pairs of x_1 entities that have k co-occurrences in their associations with x_2 entities across the entire collection of papers. Figure 20 shows two examples of co-occurrence distributions found in collections: a) $p(k,\mathbf{C}[p,r])$, the co-occurrence of references per paper pair distribution (bibliographic coupling distribution), and b) $p(k,\mathbf{C}[r,p])$, the co-occurrence of papers per reference pair distribution (co-citation distribution).

Co-occurrence distributions can be estimated from the distribution of the magnitude of the upper or lower triangle of the corresponding co-occurrence matrix (excluding the diagonal.) Co-occurrence distributions

are useful for characterizing the clustering characteristics of entities within the collection of papers. Morris (2004), for example, used bibliographic coupling distributions and co-citation distributions as metrics to evaluate clustering characteristics of papers and references in a proposed model of literature growth.



(a) (b)

Figure 20. Example of co-occurrence distributions. Left, (a), shows co-occurrence of references per paper pair distribution (bibliographic coupling distribution). Right, (b), shows co-occurrence of papers per reference pair distribution (co-citation distribution).

6.5 Clustering coefficient distributions

The clustering coefficient is a concept borrowed from complex network theory (Albert & Barabasi, 2002). When two entities have non-zero co-occurrence, define that as a co-occurrence link. Given an entity, i, its neighbor entities are the set of all entities with which it has a co-occurrence link. Define k_i as the number of neighbors of entity i and y_i as the number of links among the neighbors of entity i. This doesn't include links to entity i. The clustering coefficient c_i, of entity i, is the fraction of all possible links among the neighbors of entity i that are non-zero. That is, the number of links among the neighbors of entity i, divided by the number of possible links among those neighbors:

$$c_i(\mathbf{C}[x_1;x_2]) = \frac{y_i}{\frac{1}{2}k_i(k_i-1)} \tag{62}$$

The x_1 *co-occurrence of* x_2 *clustering coefficient* is a measure of the local connectivity in a network of x_1 entities, and is analogous to the clustering coefficient used in complex networks theory as a measure of connectedness among nodes in a network (Albert & Barabasi, 2002) Figure 21 shows a diagram of the process of calculating the co-occurrence clustering coefficient for an entity using the co-occurrence matrix.

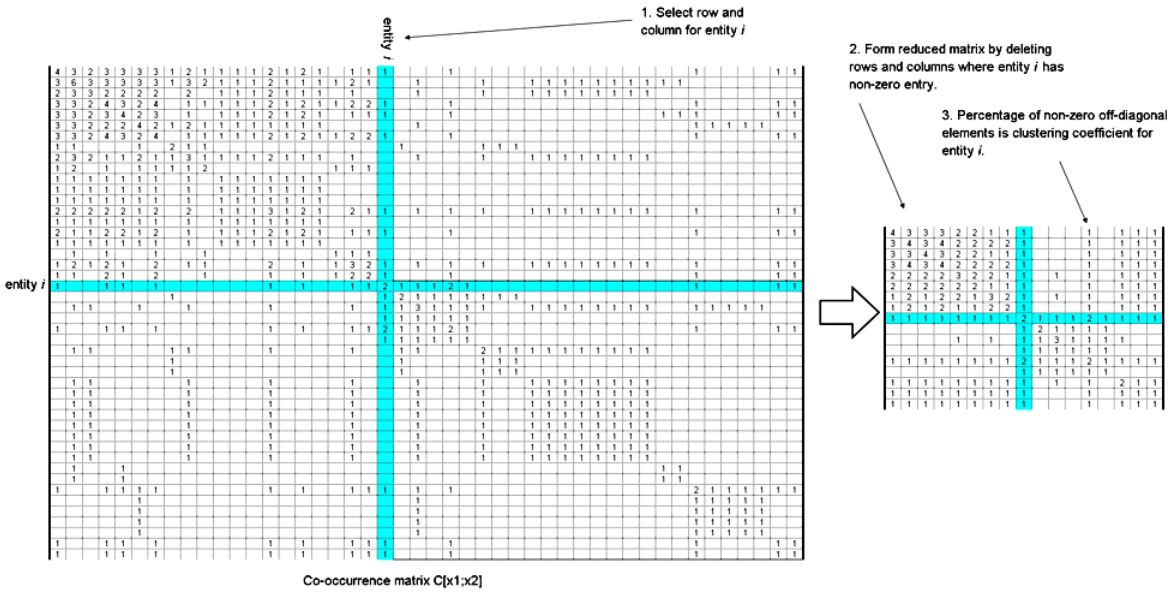


Figure 21. Diagram illustrating calculation of clustering coefficient for an entity *i* from a co-occurrence matrix.

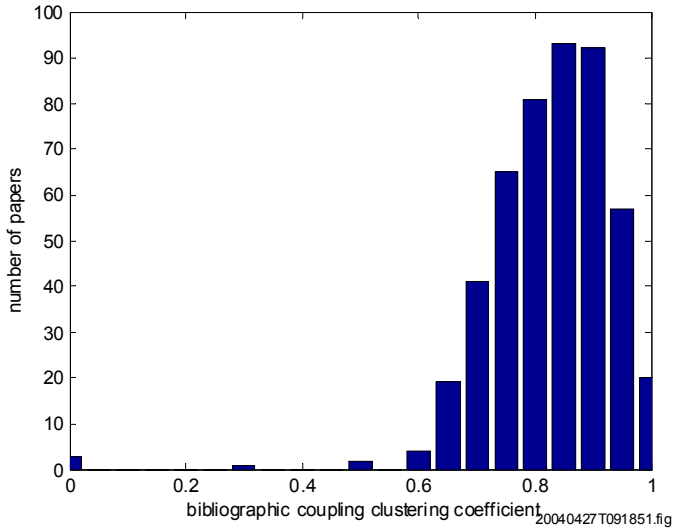


Figure 22. Example bibliographic coupling clustering coefficient distribution.

The co-occurrence clustering coefficient is a measure of the tendency of groups of entities to be locally connected. The *x_1 co-occurrence of x_2 clustering coefficient distribution*:

$$f(c, \mathbf{C}[x_1; x_2]) \tag{63}$$

is a continuous probability density distribution of the clustering coefficient, c, that ranges from zero to unity. An example of this type of distribution, shown in Figure 22, is the paper co-occurrence of references clustering coefficient distribution, $f(c,\mathbf{C}[p,r])$ (bibliographic coupling clustering coefficient distribution.) This is a measure of the tendency of local groups of papers to cover the same research topic, as measured by co-occurrence of references among papers. Morris (2004) used this distribution to compare clustering characteristics of papers from actual collections to simulations based on a proposed model of the manifestation of the emergence of scientific specialties in a collection of papers.

This chapter has presented three types of distributions that appear in collections of papers: 1) dyadic distributions, which measure the distribution of the number of associations that an entity has with entities of other entity-types, 2) co-occurrence distributions, which measure the distribution of the number of common associations that a pair of entities will have with entities from some other specific entity-type, and 3) co-occurrence clustering coefficient distributions, which measure the distribution of clustering coefficient in some co-occurrence network of like entities. These distributions are the result of growth processes in the scientific specialty associated with a collection of papers and therefore are candidates of measures of those processes.

It is beyond the scope of this report to discuss in detail the detailed characteristics of bibliometric distributions and the interpretation of their measured parameters. Discussing this topic briefly however, it has been shown in this chapter that there is a tendency of cumulating distributions to generate power laws that are characteristic of cumulative advantage processes. As such, the measured parameters of cumulating distributions may allow interpretation of the underlying cumulative advantage processes in the scientific specialty. There is already a large body of research on this topic (White & McCain, 1989) There has, however, been little research on fixed dyadic distributions, co-occurrence distributions and clustering coefficient distributions. Morris (2004) has shown that the reference per paper distribution, a fixed dyadic distribution, must be measured and modeled in order to build a realistic model of growth of paper collections. Additionally, in the same work, Morris shows that bibliographic coupling distributions and co-citation distributions, which are co-occurrence distributions, and the bibliographic coupling clustering coefficient distribution, are excellent indicators of local clustering within a collection of papers and references.

7. RECURSIVE MATRIX GROWTH

7.1 Introduction

Many of the techniques used for mapping, search, and forecasting from collections of journal papers are based on processing the entire collection of papers. There is little research that addresses the problem of dynamic updating of collections of papers.

The recursive growth equations presented in this chapter are a natural outgrowth of the matrix-based mathematical treatment of collections of journal papers introduced in this report. A dynamic expression of the growth of the collection of journal papers can have the following benefits:

- Provide insight into the dynamics of growth in the collection of papers, particularly in the growth of fixed and cumulating distributions.
- Suggest efficient methods of processing of existing paper collections.
- Suggest efficient update schemes for revising analyses results as new papers are added to a collection of papers.

The basic record in a collection of journal papers is the paper. The collection grows paper by paper in the temporal order of the publication dates of the papers. The index of papers is sequential by time. When a new paper is added, it is associated with the existing entities in the collection and additionally, new entities, e.g., new paper authors or new references, and new terms enter into the collection.

This chapter will present a recursive model of the growth of both occurrence and co-occurrence matrices as papers are added to the collection. The recursive model of matrix growth is found by examination of matrix partitions in occurrence and co-occurrence matrices as papers are added to the collection. Some applications of this recursive representation are immediately apparent. For example, the recursive model suggests methods of quickly and efficiently updating occurrence and co-occurrence matrices stored in computer memory. The recursive model also allows easy identification of static and cumulating links among the possible entity pairs within the collection of papers. Note, however that this work is preliminary and that research to explore, in depth, the possible applications of this model are out of the scope of this study.

7.2 Growth of the paper-reference matrix

It is easiest to consider the growth of an example occurrence matrix; then generalize this example to other occurrence matrices. For convenience, the paper-reference matrix will be designated **Ω**. In the matrix **Ω** the rows correspond to papers and are ordered in the sequence of publication of the papers to which they correspond. The columns correspond to references and are ordered in the sequence in which their corresponding references first appear. As shown in Figure 23, the matrix **Ω** contains a descending stair step sequence of ones from its upper left corner diagonally to its lower right corner. This sequence of ones corresponds to the initial appearance of references as papers are added to the collection. Below this diagonal sequence of ones is a roughly lower triangular region sparsely populated with ones that correspond to citations to existing references as each paper is added. Above the diagonal sequence of ones is a roughly upper triangular area of zeros.

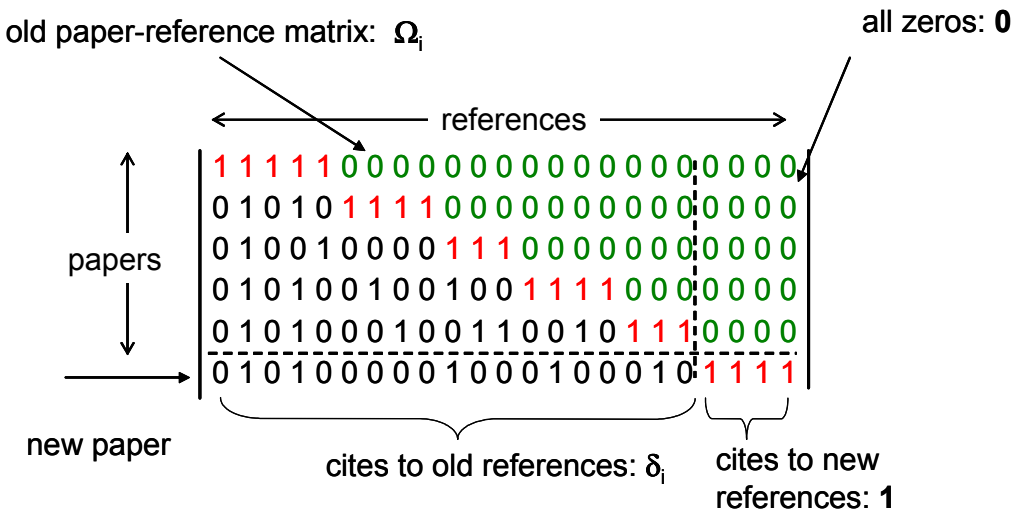


Figure 23. Diagram of the structure of a paper to reference matrix.

7.3 Dynamic growth of the paper-reference matrix

Considering the collection of journal papers dynamically, the collection grows from an initial paper by sequential addition of papers in the order in which they were published. In this sense the paper-reference matrix **Ω** grows dynamically one paper at a time. Assume i to be the number of papers, while nr_i is the number of references that have appeared in all papers up to and including paper i. Assume $\mathbf{\Omega}_i$, whose size is i by nr_i, as the paper-reference matrix after the addition of paper i, then consider the addition of paper i+1. A new row vector, i+1, is added to $\mathbf{\Omega}_i$. This vector is partitioned into a 1 by i vector $\boldsymbol{\delta}_i$ listing the paper's citations to existing references, and **1**, a 1 by nr_{i+1}-nr_i vector of ones occurring in new columns added for the new references that have appeared with paper i+1. Figure 23 shows a pictorial representation

of this addition. In the new columns, **0**, an i by nr_{i+1}-nr_i zero matrix appears. The recursive matrix equation for growth of the paper-reference equation is:

$$\mathbf{\Omega}_{i+1} = \begin{bmatrix} \mathbf{\Omega}_i & \mathbf{0} \\ \boldsymbol{\delta}_i & \mathbf{1} \end{bmatrix} \tag{64}$$

Figure 24 shows a map of a typical paper-reference matrix, where each dot shows the location of a one in the matrix. This collection of 404 papers contains 9892 citations to 6791 references. The collection was constructed using ISI's Web of Science product to find all papers that cite a seminal 1967 sociology paper by Milgrams titled "Small-World Problem."

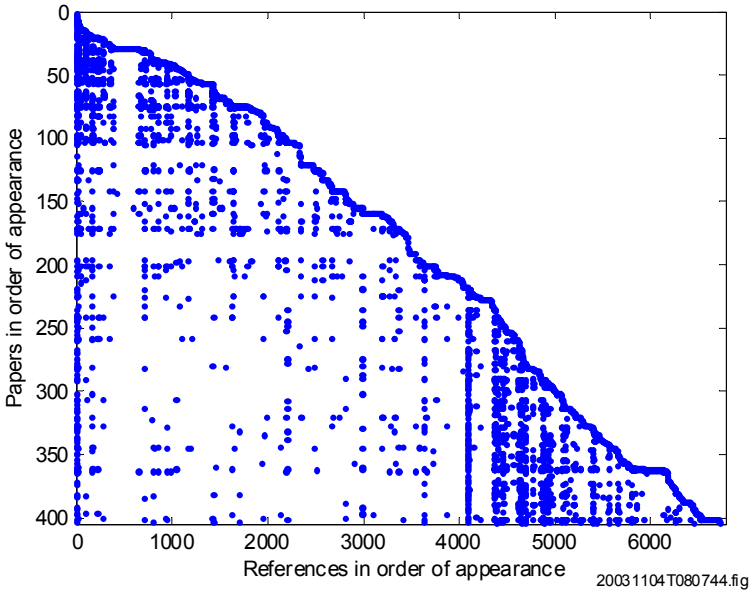


Figure 24. Example paper to reference matrix, from a collection of papers citing Milgram's 1967 "Small Worlds" paper.

7.4 Dynamic growth of the bibliographic coupling matrix

The bibliographic coupling matrix, which will be designated $\boldsymbol{\beta}$, is a symmetric matrix that lists the bibliographic coupling counts of all pairs of papers within the data collection. The diagonal of $\boldsymbol{\beta}$ contains the counts of the number of references cited in each paper. The bibliographic coupling matrix $\boldsymbol{\beta}$ can be obtained by multiplying the paper-reference matrix by its transpose:

$$\boldsymbol{\beta} = \mathbf{\Omega} \cdot \mathbf{\Omega}^T \tag{65}$$

The recursive growth equations for the bibliographic coupling matrix can be derived by substituting (64) into (65):

$$\boldsymbol{\beta}_{i+1} = \boldsymbol{\Omega}_{i+1} \cdot \boldsymbol{\Omega}_{i+1}{}^{T} = \tag{66}$$

$$\begin{bmatrix} \boldsymbol{\Omega}_i \cdot \boldsymbol{\Omega}_i{}^{T} & \boldsymbol{\Omega}_i \boldsymbol{\delta}_i{}^{T} \\ \boldsymbol{\delta}_i \boldsymbol{\Omega}_i{}^{T} & (\boldsymbol{\delta}_i \,|\, \mathbf{1})(\boldsymbol{\delta}_i \,|\, \mathbf{1})^{T} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_i & \boldsymbol{\Omega}_i \boldsymbol{\delta}_i{}^{T} \\ \boldsymbol{\delta}_i \boldsymbol{\Omega}_i{}^{T} & m_{i+1} \end{bmatrix}$$

where $(\boldsymbol{\delta}_i|\mathbf{1})$ is the bottom row of $\boldsymbol{\beta}_{i+1}$, i.e., the concatenation of $\boldsymbol{\delta}_i$ and $\mathbf{1}$. Also, m_{i+1} is the number of references cited by paper i+1. Figure 25 shows a pictorial representation of a typical bibliographic coupling matrix with the partitions in Equation (66) identified. It is easy to see from Equation (66) and Figure 25 that bibliographic coupling counts between pairs of papers are static, and do not change as more papers are added to the collection.

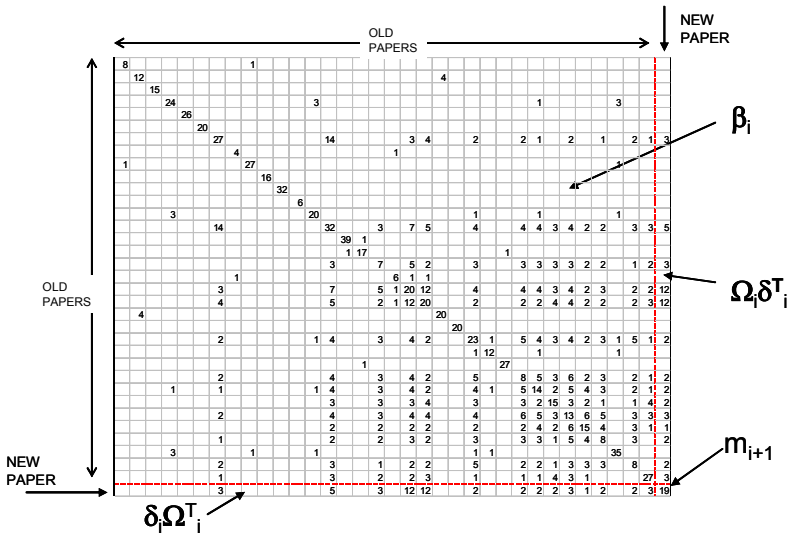


Figure 25. Diagram of a bibliographic coupling matrix.

7.5 Dynamic growth of the co-citation matrix

The co-citation matrix, designated $\boldsymbol{\Gamma}$, is a symmetric nr by nr matrix that lists the co-citation counts of all pairs of references within the data collection. The diagonal of $\boldsymbol{\Gamma}$ contains the counts of the number of papers that cite each reference. The co-citation matrix $\boldsymbol{\Gamma}$ can be obtained by multiplying the transpose of the paper-reference matrix by itself:

$$\mathbf{\Gamma} = \mathbf{\Omega}^T \cdot \mathbf{\Omega} \tag{67}$$

The recursive growth equations for the co-citation matrix can be derived by substituting (64) into (67):

$$\mathbf{\Gamma}_{i+1} = \mathbf{\Omega}_{i+1}{}^T \cdot \mathbf{\Omega}_{i+1} =$$

$$\begin{bmatrix} \mathbf{\Omega}_i{}^T \mathbf{\Omega}_i + \boldsymbol{\delta}_i{}^T \boldsymbol{\delta}_i & \boldsymbol{\delta}_i{}^T \mathbf{1} \\ \mathbf{1}^T \boldsymbol{\delta}_i & \mathbf{1}^T \mathbf{1} \end{bmatrix} = \begin{bmatrix} \mathbf{\Gamma}_i + \boldsymbol{\delta}_i{}^T \boldsymbol{\delta}_i & \boldsymbol{\delta}_i{}^T \mathbf{1} \\ \mathbf{1}^T \boldsymbol{\delta}_i & \mathbf{1}^T \mathbf{1} \end{bmatrix} \tag{68}$$

Figure 26 shows a pictorial representation of a typical co-citation matrix with the partitions in (68) identified. It is easy to see that the co-citation count between two references is not static, but can be increased with the addition of each new paper to the collection.

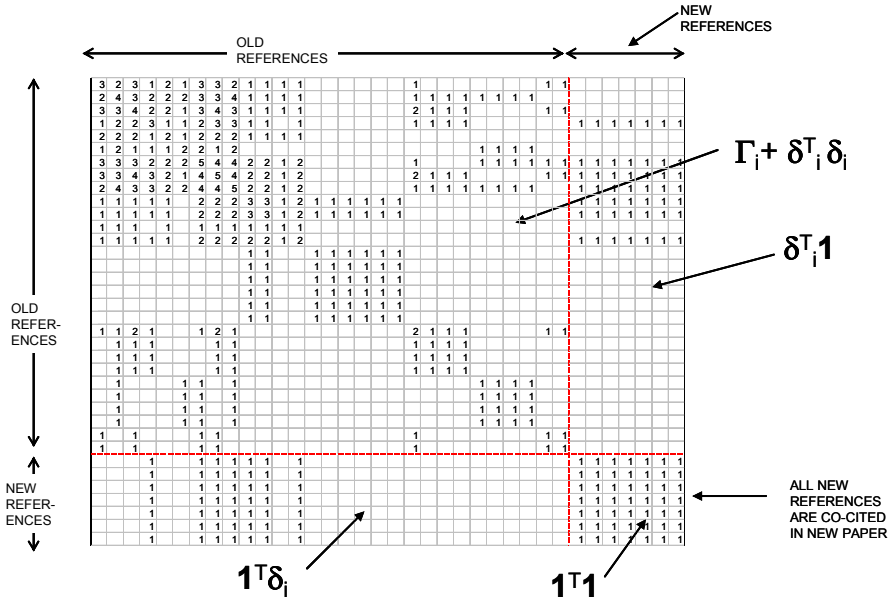


Figure 26. Diagram of a co-citation matrix.

7.6 General growth of occurrence and co-occurrence matrices

It is easy to generalize the example of the paper-reference matrix to other occurrence and co-occurrence matrices. Given a pair of entity-types, we define the dependent entity-type as the entity-type that depends on the creation of the first entity-type to be created. References, paper authors, paper journals, paper authors and terms within a collection of papers are dependent on papers. Reference authors and reference journals are dependent on references. For a generalization of the recursive matrix equations for occurrence

and co-occurrence matrices assume that the independent entity-type corresponds to rows of an occurrence matrix and that the dependent entity-type is placed on the columns of the matrix. Given entity-types x_1 and x_2, where x_2 is the dependent entity-type, the recursive matrix equation for the occurrence matrix is:

$$\mathbf{\Omega}_{i+1} = \begin{bmatrix} \mathbf{\Omega}_i & \mathbf{0} \\ \boldsymbol{\delta}_i & \boldsymbol{\eta} \end{bmatrix} \tag{69}$$

Where:

- $\mathbf{\Omega}$ is the occurrence matrix $\mathbf{O}[x_1;x_2]$
- $\boldsymbol{\delta}_i$ is an $nx_1(i+1)$-$nx_1(i)$ by $nx_2(i)$ matrix containing associations between the new x_1 entities and the old x_2 entities. When x_1 is papers then $nx_1(i+1)$-$nx_1(i)$ is always unity because papers are added one at a time. When x_1 is references then $nx_1(i+1)$-$nx_1(i)$ can be greater than unity as multiple references can be added with each paper. Note that the paper index is placed between parentheses in the notation to avoid confusion with entity-type index subscripts.
- $\boldsymbol{\eta}$ is an $nx_1(i+1)$-$nx_1(i)$ by $nx_2(i+1)$- $nx_2(i)$ matrix containing associations between the new x_1 entities and the new x_2 entities.

The co-occurrence matrix for the row entity-type is:

$$\boldsymbol{\beta}_{i+1} = \begin{bmatrix} \boldsymbol{\beta}_i & \mathbf{\Omega}_i \boldsymbol{\delta}_i^{\,T} \\ \boldsymbol{\delta}_i \mathbf{\Omega}_i^{\,T} & \boldsymbol{\eta}\boldsymbol{\eta}^T \end{bmatrix} \tag{70}$$

Where $\boldsymbol{\beta}$ is the co-occurrence matrix $\mathbf{C}[x_1;x_2]$. The co-occurrence matrix for the column entity-type is:

$$\mathbf{\Gamma}_{i+1} = \begin{bmatrix} \mathbf{\Gamma}_i + \boldsymbol{\delta}_i^{\,T} \boldsymbol{\delta}_i & \boldsymbol{\delta}_i^{\,T} \boldsymbol{\eta} \\ \boldsymbol{\eta}^T \boldsymbol{\delta}_i & \boldsymbol{\eta}^T \boldsymbol{\eta} \end{bmatrix} \tag{71}$$

Where $\mathbf{\Gamma}$ is the co-occurrence matrix $\mathbf{C}[x_2;x_1]$. Note that the recursive matrix equations of Equation (70) and Equation (71) can easily be generalized for using link weight functions other than matrix multiplication. See Chapter 3.2.

The recursive growth model introduced in this chapter may have several applications. The recursive equations themselves suggest methods of saving memory and computation when updating matrices in the computer memory. While this may be useful for very large collections of papers, it has not been found

necessary in practical experience. Most topics of interest are covered sufficiently by just two or three thousand papers, and such collections do not present unmanageable computational difficulties. This is particularly true when using MATLAB when computation can be "vectorized" and MATLAB's sparse matrix algorithms can be applied. Two computational problems are encountered for large collections: 1) calculation of similarities, and 2) hierarchical clustering. The recursive approach presented here may allow efficient recursive methods of attacking these two computational problems.

The methods are also possibly applicable to building viable growth models for complex networks. There are several key complex network growth models being investigated currently (Albert & Barabasi, 2002; Dorogovtsev & Mendes, 2002). The most important of these are preferential attachment models related to the Yule model, of which the popular Barabasi-Albert model (Albert & Barabasi, 2002) is a special case. All of these models are single entity-type models that are very limited in scope. The dynamic equations introduced here should allow building complex multiple entity-type models of complex networks, a more realistic approach. The ability to model multiple entity-types and, in particular, the ability of these recursive matrix equations to discriminate between fixed and cumulating links promises to add a new dimension to growth models of complex networks. This will enable the dynamic modeling and simulation of dyadic, co-occurrence, and clustering coefficient distributions. This recursive formulation of growth of paper collections also holds promise for analysis of causal relations in networks, through the distinction between fixed and cumulating relations, which will allow extraction of practical and useful analysis of complex multiple entity-type networks.

8. GRAPH THEORETIC MATRICES

The term 'graph theoretic' is used to denote analysis methods based on direct citations from paper to paper within the collection. This type of analysis is often used to trace flow of information and seminal ideas from paper to paper (Garfield et al., 2003). A model of a collection of papers as a graph connected by citations is the key basis for a collection of knowledge mapping techniques introduced by Small (1997). Small's method of clustering papers for analysis is based on deriving inter-paper similarities on a combination of 4 different types of graph theoretic links in the network of papers. These graph theoretic link types are 1) direct citation, 2) longitudinal coupling, 3) bibliographic coupling and 4) co-citation. This chapter will discuss the derivation of these four types of links in the context of the matrix formulation and notation introduced in this report.

As noted in Chapter 2.3 and Chapter 4.3, it is necessary to match references and papers to the same physical entity in order to find the adjacency matrix of papers linked by their citations. The paper to cited paper adjacency matrix, $\mathbf{O}[p;cp]$, is calculated from the paper to reference matrix $\mathbf{O}[p;r]$ and the reference to cited paper equivalency matrix $\mathbf{A}[r;cp]$:

$$\mathbf{O}[p;cp] = \mathbf{O}[p;r] \cdot \mathbf{A}[r;cp] \tag{72}$$

Matrices for each of the four types of graph theoretic links can easily be calculated through matrix arithmetic operations on the paper to cited paper adjacency matrix. The diagram of Figure 27 shows four types of links as they occur on the paper graph. Each diagram shows a pair of papers, *i* and *j*, and a possible third paper *k*. and the links among them that comprise the graph theoretic link.

8.1 Direct citation

A direct citation link between two papers occurs if one paper cites the other:

$$dc_{ij} = \begin{cases} 1, & \text{if } p_i \text{ cites } p_j \text{ or } p_j \text{ cites } p_i \\ 0, & \text{otherwise} \end{cases} \tag{73}$$

The matrix **DC** is a symmetric binary matrix that lists all the direct citation links in the paper graph. It can be calculated by adding the adjacency matrix to its transpose:

$$\mathbf{DC} = \mathbf{O}[p;cp] + \mathbf{O}[cp;p] \tag{74}$$

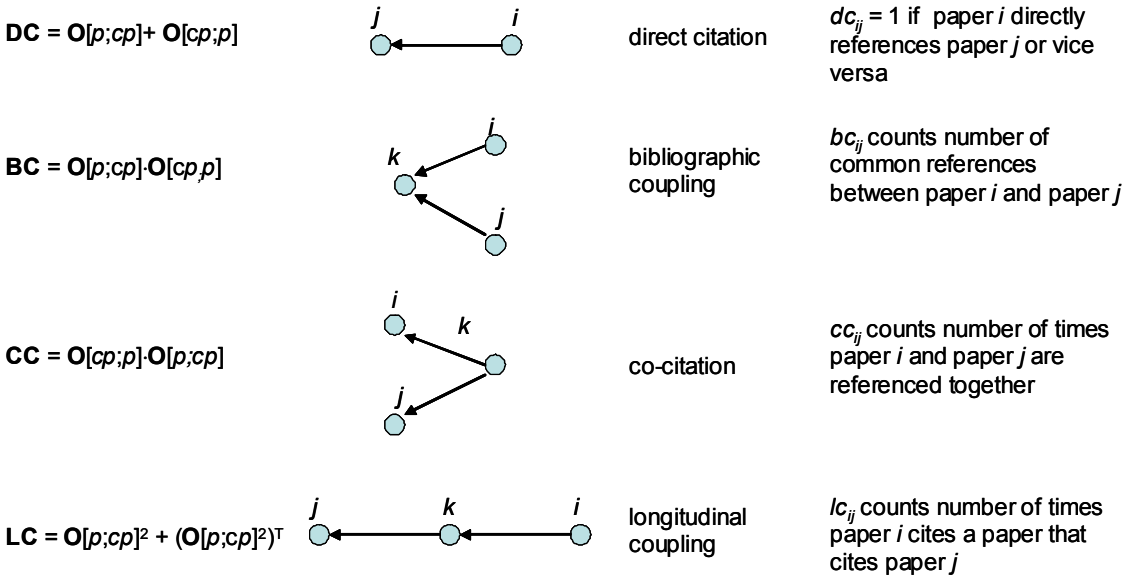


Figure 27. Diagram of paper to paper links based on graph theoretic paths.

.

8.2 Longitudinal coupling

A longitudinal coupling link between a pair of papers occurs if one paper cites a third paper that cites the other paper of the pair:

$$lc_{ij}(k) = \begin{cases} 1, & \text{if } p_i \text{ cites } p_k \text{ AND } p_k \text{ cites } p_j \\ 1, & \text{if } p_j \text{ cites } p_k \text{ AND } p_k \text{ cites } p_i \\ 0, & \text{otherwise} \end{cases} \tag{75}$$

The longitudinal coupling count between a pair of papers is the number of longitudinal coupling links between them:

$$lc_{ij} = \sum_k lc_{ij}(k) \tag{76}$$

The matrix **LC** is a symmetric non-binary matrix that lists all the longitudinal coupling links in the paper graph. It can be calculated from the square of the adjacency matrix added to the transpose of the square of the adjacency matrix:

$$\mathbf{LC} = \mathbf{O}[p;cp]^2 + \left(\mathbf{O}[p;cp]^2\right)^T \tag{77}$$

8.3 Co-citation

Co-citation of cited papers is completely analogous to co-citation of references as discussed in Chapter 5.3. A co-citation link occurs between a pair of papers if both papers are cited by a common third paper:

$$cc_{ij}(k) = \begin{cases} 1, & \text{if } p_k \text{ cites } p_i \text{ AND } p_k \text{ cites } p_j \\ 0, & \text{otherwise} \end{cases} \tag{78}$$

The co-citation count between a pair of papers is the number of co-citation links between them:

$$cc_{ij} = \sum_k cc_{ij}(k) \tag{79}$$

The matrix **CC** is a symmetric non-binary matrix that lists all the co-citation counts in the paper graph. It can be calculated by pre-multiplying the graph adjacency matrix by its transpose:

$$\mathbf{CC} = \mathbf{O}[cp;p] \cdot \mathbf{O}[p;cp] \tag{80}$$

8.4 Bibliographic coupling

Bibliographic coupling between papers in a citation adjacency matrix is completely analogous to bibliographic coupling of papers citing common references as discussed in Chapter 5.3. A bibliographic coupling link occurs between a pair of papers if both of them cite a common third paper:

$$bc_{ij}(k) = \begin{cases} 1, & \text{if } p_i \text{ cites } p_k \text{ AND } p_j \text{ cites } p_i \\ 0, & \text{otherwise} \end{cases} \tag{81}$$

The bibliographic coupling count between a pair of papers is the number of co-citation links between them:

$$bc_{ij} = \sum_k bc_{ij}(k) \tag{82}$$

The matrix **BC** is a symmetric non-binary matrix that lists all the bibliographic coupling counts in the paper graph. It can be calculated by post-multiplying the graph adjacency matrix by its transpose:

$$\mathbf{BC} = \mathbf{O}[p; cp] \cdot \mathbf{O}[cp; p] \tag{83}$$

The four matrices of graph theoretic linkages described in this chapter can be used for both calculation of Small's similarity measure and for analysis of information flow among papers. When calculating Small's similarity measure, it is necessary to fuse the weights from the four types of links into a single similarity measure. Chapter 9.3 will discuss the fusion technique introduced by Small and generalize that technique to allow arbitrary weighting of the four different link types in the similarity calculation.

9. SIMILARITY

9.1 Calculation of similarity

A *similarity value* is a measure of the similarity between a pair of entities of the same entity-type and is used for analysis of the relations between entities in the collection of papers. Similarity values range from zero for no similarity to unity for identical entities. Converting raw co-occurrence values to similarities before analysis is usually advantageous because this tends to suppress artifacts caused by dominant entities with large numbers of links.

Similarities are usually calculated by normalizing co-occurrence counts. Many normalization schemes exist. Salton (1989) gives a review of several similarity formulas, while Jones and Furnas (1987) discuss some of the advantages of each. The cosine coefficient, Jaccard coefficient, and the dice coefficient will be reviewed here. To simplify notation, use the following simplified variables:

- $s_{ij}[x_1;x_2] \rightarrow s_{ij}$: is the similarity between entity i and entity j of the x_1 entity-type based on co-occurrence with the x_2 entity-type
- $c_{ij}[x_1;x_2] \rightarrow c_{ij}$: is an element in the co-occurrence matrix and is the co-occurrence count between entity i and entity j of the x_1 entity-type based on co-occurrence with the x_2 entity-type
- $c_{ii}[x_1;x_2] \rightarrow c_{ii}$: is from the diagonal of the co-occurrence matrix and is the occurrence count of entity i of the x_1 entity-type, that is, the number of times entity i is associated with an x_2 entity-type.

The similarities are calculated from elements in the co-occurrence matrix. As shown in Figure 28, similarities between entity i and entity j are calculated using three elements from the co-occurrence matrix: 1) the co-occurrence count c_{ij}, 2) the occurrence count c_{ii} and 3) the occurrence count c_{jj}. Note that for similarity to properly range from zero to unity, c_{ij}, c_{ii}, and c_{jj} must all be greater than zero and the following constraints must be satisfied:

$$2c_{ij} < c_{ii} + c_{jj} \tag{84}$$

$$c_{ij}^{\,2} < c_{ii}c_{jj} \tag{85}$$

All of the link weight functions discussed in Chapter 3.2 produce co-occurrence matrices whose elements satisfy these constraints, these functions include matrix multiplication, the overlap function, and the inverse Minkowski function.

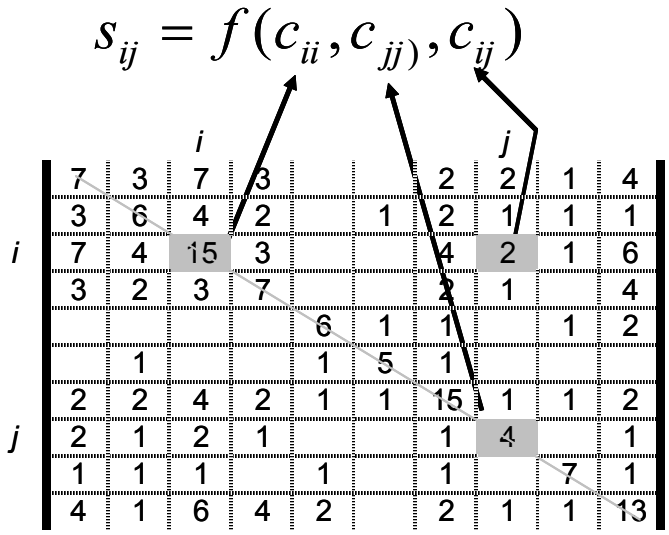


Figure 28. Diagram showing the co-occurrence matrix elements used for computation of similarity. Similarity s_{ij} is computed from element c_{ij} and diagonal elements c_{ii} and c_{jj}.

The cosine coefficient is given by:

$$s_{ij} = \frac{c_{ij}}{\sqrt{c_{ii} \cdot c_{jj}}} \tag{86}$$

The dice coefficient is given by:

$$s_{ij} = \frac{2c_{ij}}{c_{ii} + c_{jj}} \tag{87}$$

The Jaccard coefficient is given by:

$$s_{ij} = \frac{c_{ij}}{c_{ii} + c_{jj} - c_{ij}} \tag{88}$$

The three similarity formulas shown above all produce maximum similarity equal to unity when $c_{ii} = c_{jj} = c_{ij}$ and reach minimum at zero when c_{ij} is equal to zero. For purpose of calculation, the similarities cannot be directly computed using matrix multiplication. However, it is possible using sparse matrix techniques, such as those available in MATLAB, to vectorize the computations for very efficient calculation.

9.2 Fusion of similarity from co-occurrence of multiple entity-types

Similarities can be combined to give a similarity value derived from more than one co-occurrence value. A straightforward method of similarity fusion is simply:

$$s_{ij}[x_1; all(x_k)] = \sum_k a_k s_{ij}[x_1; x_k] \quad \text{where } \sum_k a_k = 1 \text{ and } 0 \le a_k \le 1 \tag{89}$$

As an example consider:

$$s_{ij}[p; ap, r] = a \cdot s_{ij}[p; ap] + (1-a) \cdot s_{ij}[p; r] \tag{90}$$

Equation (90) fuses similarity from bibliographic coupling with similarity derived from co-occurrence of paper authors, papers linked by common paper authors, a measure used by Asnake (2003) to improve clustering of papers into research fronts. Some other fused multiple entity-type similarities are the dice coefficient:

$$s_{ij}[x_1; all(x_k)] = \frac{2\sum_k w_k c_{ij}[x_1; x_k]}{\sum_k w_k \left(c_{ii}[x_1; x_k] + c_{jj}[x_1; x_k]\right)} \tag{91}$$

and the cosine coefficient:

$$s_{ij}[x_1; all(x_k)] = \frac{\sum_k w_k c_{ij}[x_1; x_k]}{\sqrt{\sum_k (w_k c_{ii}[x_1; x_k]) \cdot \sum_k (w_k c_{jj}[x_1; x_k])}} \tag{92}$$

The weights, $w_1, w_2, w_3, \ldots, w_k$ in the two equations above are greater than or equal to zero and are used to adjust the relative emphasis on each of the similarity measures in the calculation. A threshold on occurrences can be used to reduce the resulting similarity if occurrences of a particular type fall below

some threshold. As an example Equation (92) for the fusion cosine formula can be modified to include occurrence thresholds for each measure:

$$s_{ij}[x_1; all(x_k)] = \frac{\sum_k w_k c_{ij}[x_1; x_k]}{\sqrt{\sum_k w_k \min(\delta_k, c_{ii}[x_1; x_k]) \cdot w_k \sum_k \min(\delta_k, c_{jj}[x_1; x_k])}}, \tag{93}$$

where δ_1, δ_2, … δ_k are the occurrence thresholds for each measure. This similarity fusion formula will be used to generalize Small's similarity, discussed in the next section.

9.3 Small's similarity

Small's similarity (Small, 1997) is used for clustering papers using graph theoretic properties for mapping, and is a weighted, normalized sum of direct citations, bibliographic coupling, co-citations, and longitudinal coupling. This similarity measure was introduced as a method of calculating similarities to map large collections of papers. It is based on the graph theoretic relations discussed in Chapter 8 using the adjacency matrix **O**[*p*;*cp*] defined in Equation (72). The equation for Small's similarity is originally defined as:

$$s_{ij} = \frac{2dc_{ij} + bc_{ij} + cc_{ij} + lc_{ij}}{\sqrt{(1+n_i)(1+n_j)}} \tag{94}$$

where n_i is defined as:

$$n_i = nbc_i + ncc_i + nlc_i \tag{95}$$

$$n_j = nbc_j + ncc_j + nlc_j \tag{96}$$

where nbc_i, ncc_i, nlc_i, are the total number of bibliographic coupling links, co-citation links, and longitudinal coupling links connecting to paper *i* respectively. Small's similarity is an empirically derived similarity that is a compromise measure that relates a paper to the papers it cites, the papers that cite it, the papers that cite the same papers it does, and finally, to the papers that it is cited together with. Assuming that papers covering the same topic will tend to be linked in several different ways, Small's similarity is effective as a basis for clustering papers because the fusion of four different similarities tends to reduce noise in the resulting measured similarity. This makes the clustering more robust and helps to successfully classify more of the marginally related papers than would otherwise be possible.

Note, however that Small's measure can confuse the symbolic representation (a concept to be defined and discussed in Chapter 11) of groups of papers clustered using that measure. There are 6 possible relations between two papers A and B that can contribute to Small's similarity:

1. A uses B
2. A is used by B
3. A and B use C
4. A and B are both used by C
5. A used C which used B
6. A is used by C which is used by B

The relations above, if used singly, will produce groups that have some definite symbolic representation, such as groups of papers that use the same references, item 1 in the list above. See Chapter 11.1 for a discussion of symbolic representation of groups of entities. Groups formed by clustering using Small's similarity are ambiguously related in this sense, at best they can be described as groups of papers that are somehow related by what they use or how they are used.

To generalize Small's similarity, we will use the general similarity fusion formula as introduced in Chapter 9.2. Define the following matrices, taken from Chapter 8:

$$\mathbf{C}[p;x_1] = \mathbf{DC} \qquad \text{Direct citation} \tag{97}$$

$$\mathbf{C}[p;x_2] = \mathbf{BC} \qquad \text{Bibliographic coupling} \tag{98}$$

$$\mathbf{C}[p;x_3] = \mathbf{CC} \qquad \text{Co-citation} \tag{99}$$

$$\mathbf{C}[p;x_4] = \mathbf{LC} \qquad \text{Longitudinal coupling} \tag{100}$$

Variables x_1, x_2, x_3, and x_4 are dummy variables for indexing purposes. Now, using the similarity fusion formula with minimum threshold given by Equation (93), Small's similarity in general form is:

$$s[p;(x_1,x_2,x_3,x_4)]_{ij} = \frac{\sum_{k=1}^{4} w_k c_{ij}[p;x_k]}{\sqrt{\sum_{k=1}^{4} w_k \min(\delta_k, c_{ii}[p;x_k]) \cdot \sum_{k=1}^{4} w_k \min(\delta_k, c_{jj}[p;x_k])}} \tag{101}$$

where w_1, w_2, w_3, w_4 are weights used to adjust the relative emphasis on each of the measures: direct citation, bibliographic coupling, co-citations, and longitudinal coupling respectively. The thresholds δ_1, δ_2, δ_3, δ_4 are used to adjust the importance of the missing measure. Comparing Equation (101) to Equation (94), for Small's similarity as originally proposed the following weights and thresholds are used.

- $w_1 = 2$ $\delta_1 = ½$
- $w_2 = 1$ $\delta_2 = 0$
- $w_3 = 1$ $\delta_3 = 0$
- $w_4 = 1$ $\delta_4 = 0$

This results in direct citation linkage receiving twice as much emphasis as the other measures and direct citation receiving a threshold of ½ while other measures have no threshold.

As expressed in this chapter the similarities can be systematically expressed and used in terms of the co-occurrence matrices in the collection of papers. Furthermore the similarity fusion formulas shown here are general and easily adaptable to fuse similarities for any primary entity-type in relation to multiple relative entity-types. Using programming languages adapted to matrix arithmetic, such as MATLAB, the calculations are easily vectorized and adapted to existing fast and efficient sparse matrix functions.

10. ENTITY FEATURE VECTORS

10.1 Introduction

In the pattern recognition sense, a *feature* is a measurable observable associated with an entity that can be used to characterize an entity for purposes of clustering, mapping, and other statistical techniques. A full review of features and their use in pattern recognition can be found in Duda and Hart (2001). A *feature vector* is a vector where each element holds a feature. Usually, feature vectors are considered as coordinates in some multi-dimensional *feature space*. Given the feature vectors for a collection of entities, many techniques, such as c-means clustering or multidimensional scaling, can be applied to classify, compare and map the entities for analysis.

10.2 Types of feature vectors

Using the mathematical treatment proposed in this report it is possible to define two types of feature vectors for entities in collections of journal papers: 1) occurrence feature vectors and, 2) co-occurrence feature vectors. The occurrence vector shows the pattern of associations that an entity has with entities of an unlike entity-type, while a co-occurrence vector shows the pattern of co-occurrences that an entity has with like entities. Assume a pair of entity-types described by an occurrence matrix. From the occurrence matrix two co-occurrence matrices can be formed. Given primary entity-type, x_1, and relative entity-type, x_2, assume that the i^{th} entity of entity-type x_1, is the entity of interest, Two feature vectors can be formed for entity i that describe its relation to the relative x_2 entities:

- $\mathbf{O}_i[x_1; x_2]$: an occurrence feature vector listing the number of times each x_2 entity is associated with x_1 entity i. This corresponds to row i in the occurrence matrix $\mathbf{O}[x_1; x_2]$.
- $\mathbf{C}_i[x_1; x_2]$: a co-occurrence feature vector listing the number of times each x_1 entity co-occurs with entity i in their association with x_2 entities. This corresponds to row i (and column i since the matrix is symmetric) in the co-occurrence matrix $\mathbf{C}[x_1; x_2]$.

The length of the occurrence feature vector is the number of relative entities nx_2. The length of the co-occurrence feature vector is the number of primary entities nx_1. Figure 29 shows a diagram of a paper author to reference author occurrence matrix, $\mathbf{O}[ap;ar]$, and the associated paper author co-occurrence matrix, $\mathbf{C}[ap;ar]$. A paper author i is highlighted and its occurrence feature vector and co-occurrence

feature vector are shown. Each entity in a paper collection has *NE*-1 occurrence feature vectors and *NE*-1 co-occurrence features, where *NE* is the number of entity-types in the paper collection.

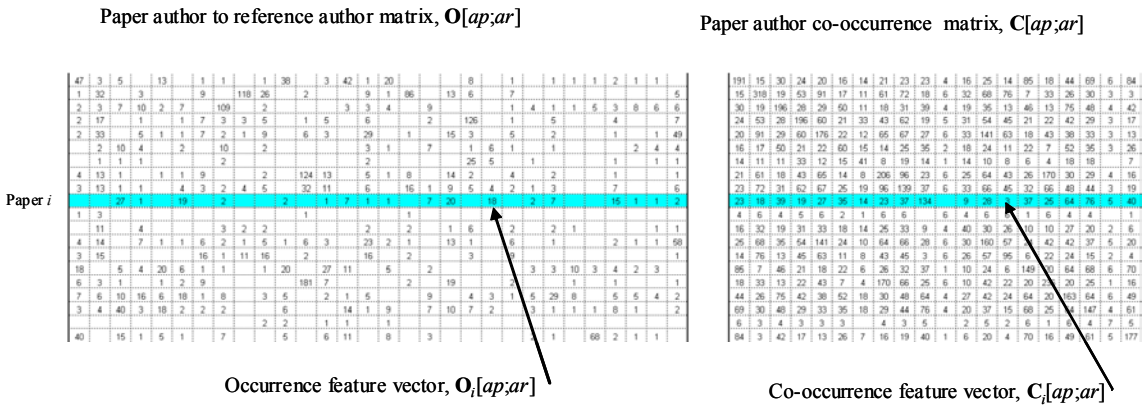


Figure 29. Example of feature vectors. Given papers as the primary entity-type and references as the relative entity-type, the occurrence feature vector for paper *i* is row *i* from the paper to reference matrix **O**[*p*;*r*], while the co-occurrence feature vector for paper *i* is row *i* (or column *i*) from the co-occurrence matrix **C**[*p*;*r*].

The occurrence vector, $\mathbf{O}_i[x_1; x_2]$, associated with x_1 entity *i*, describes the set of x_2 entities associated with x_1 entity *i* and serves as a *characterizing pattern*, that is, a pattern of associations that helps to characterize and classify x_1 entity *i.* For example, the vector $\mathbf{O}_i[ar; ap]$, listing the paper authors citing a reference author *i*, characterizes reference author *i* by the pattern of authors that read and use his or her work. Table 3 shows a list of different types of occurrence feature vectors with their associated characterizing patterns.

Table 3. Examples of occurrence feature vectors for entities in a collection of papers.

Primary entity-type x_1	**Relative entity-type x_2**	**Feature vector for entity *i***	**Characterizing pattern**
paper	reference	$\mathbf{O}_i[p;r]$	a) The concept symbols used by a paper (Small, 1978). b) the knowledge sources used by a paper.
reference	paper	$\mathbf{O}_i[r;p]$	The papers using a reference as a concept symbol.
paper author	paper	$\mathbf{O}_i[ap;p]$	A paper author's oeuvre
paper author	reference author	$\mathbf{O}_i[ap;ar]$	The reference authors whose work a paper author reads and uses. An author's *identity* (White, 2001).
reference author	paper author	$\mathbf{O}_i[ar;ap]$	The paper authors that read and use a reference author's work.
paper journal	reference journal	$\mathbf{O}_i[jp;jr]$	The reference journals holding source knowledge used by papers in a paper journal
reference journal	paper journal	$\mathbf{O}_i[jr;jp]$	The paper journals whose papers draw knowledge from a reference journal
paper	terms	$\mathbf{O}_i[p;t]$	A paper's research vocabulary

The co-occurrence vector, $\mathbf{C}_i[x_1; x_2]$, associated with x_1, gives the set of x_1 entities that are associated with the same x_2 entities as x_1 entity i. Similar to occurrence feature vectors, co-occurrence feature vectors may serve as some specific characterizing pattern. For example, the vector $\mathbf{C}_i[r;p]$ characterizes reference i by the list of the references that are co-cited with reference i. Table 4 shows a list of primary entity-type to relative entity-type pairs and characterizing patterns that can be derived from the associated co-occurrence feature vectors.

Table 4. Examples of co-occurrence feature vectors for entities in a collection of papers.

Primary entity-type x_1	**Relative entity-type x_2**	**Feature vector for entity i**	**Characterizing pattern**
paper	reference	$\mathbf{C}_i[p;r]$	The papers that use the same concept symbols as paper *i.* (Papers covering the same topic as paper *i.*)
reference	paper	$\mathbf{C}_i[r;p]$	The references being used by the same papers the use reference *i*. (Exemplar references for the same Kuhnian paradigm as reference *i.*)
paper author	paper	$\mathbf{C}_i[ap;p]$	The collaborators of paper author *i.*
paper author	reference author	$\mathbf{C}_i[ap;ar]$	The paper authors using the same knowledge sources as paper author *i.* Paper author *i*'s *invisible college*.
reference author	paper author	$\mathbf{C}_i[ar;ap]$	The reference authors used as knowledge sources by the same paper authors as reference author *i*. The *image* of reference author *i*. (White, 2001)
paper journal	reference journal	$\mathbf{C}_i[jp;jr]$	The paper journals using the same sources of knowledge as paper journal *i*.
reference journal	paper journal	$\mathbf{C}_i[jr;jp]$	The reference journals (sources of knowledge) being used by the same paper journals as reference journal *i*.
paper	terms	$\mathbf{C}_i[p;t]$	Papers using the same research vocabulary as paper *i*. (Papers covering the same topic as reference journal *i*.)

Occurrence and co-occurrence feature vectors, together with the similarities discussed in Chapter 9, provide measurements of inter-entity relations that can be used to map the structure of the entity groups in a collection of papers. Chapter 11 will discuss the seriation and clustering methods that are used to map and group entities in the collection of papers.

11. SERIATION, CLUSTERING AND ENTITY GROUPS

11.1 Introduction

Seriation is a method to order lists of entities according to some criteria. *Clustering* is the process of finding groups of entities according to some criteria. Both methods are important for uncovering structure among the entities within a collection of papers. Most seriation and clustering routines use inter-entity similarities derived from co-occurrence matrices as input to their algorithms. For example, seriation based on the traveling salesman problem (TSP) attempts to place entities that are most similar to each other in an ordering by maximizing the sum of similarities of adjacent entity pairs in the ordering (Bar-Joseph, Gifford, & Jaakola, 2001). Hierarchical agglomerative clustering, the most commonly used clustering technique, produces clusters by an iterative bottom-up fusing of the two most similar entities or clusters in the collection (Gordon, 1999). For seriation and clustering methods based on comparing distances between entities, distances can be calculated from similarities by subtracting them from unity or by other well-known methods (Jones & Furnas, 1987).

When groups of primary entities are found based on co-occurrence with a relative entity, it is important to consider what such groups represent. Groups of entities clustered on co-occurrence share a *common characteristic*. Given the common characteristic of such groups, they can be assigned a *symbolic representation*. For example, groups of paper authors that are formed by clustering on co-occurrence of papers have a common characteristic of "common papers." From this characteristic it can be inferred that a group of authors clustered this way have co-authored one or more papers together. The symbolic representation for such groups then would be *collaboration groups*. Figure 30 shows an entity relationship diagram listing several useful symbolic representations for groups of primary entities formed by clustering on co-occurrence with different relative entities. A primary entity-type may yield several symbolic representations depending on the relative entity-type. For example, groups of papers formed by clustering on co-occurrence with paper authors share a common characteristic of "common authors" and can be assigned a symbolic representation of "collaboration group oeuvres." However, groups of papers can also be formed by clustering on co-occurrence of references. A group of such papers would have a common characteristic of "common references." A group of papers using the same references are using the same concept symbols and it follows that they are reporting on the same research topic. The symbolic representation of such a group could be "research front," meaning a group of papers covering a common topic or Kuhnian puzzle.

Aside from using similarities, it is possible to work with feature vectors as input to vector based clustering methods, such as c-means clustering, to produce clusters. These methods are based on iteratively adjusting the locations of cluster centers in the feature space. Each entity is assigned to the cluster whose center is closest to its feature vector.

The discussion in this chapter will center on the effects of clustering on the occurrence and co-occurrence matrices within the collection of papers. Rather than reviewing extensively the mechanics of seriation and clustering, the discussion will focus on the results of such methods and how they are related to the permutation and shading of occurrence matrices in a collection of papers. This will aid in understanding the process of clustering and seriation as the ordering of entities and permutation of occurrence matrices in a way that reveals structure in the relation of entities in the collection of papers.

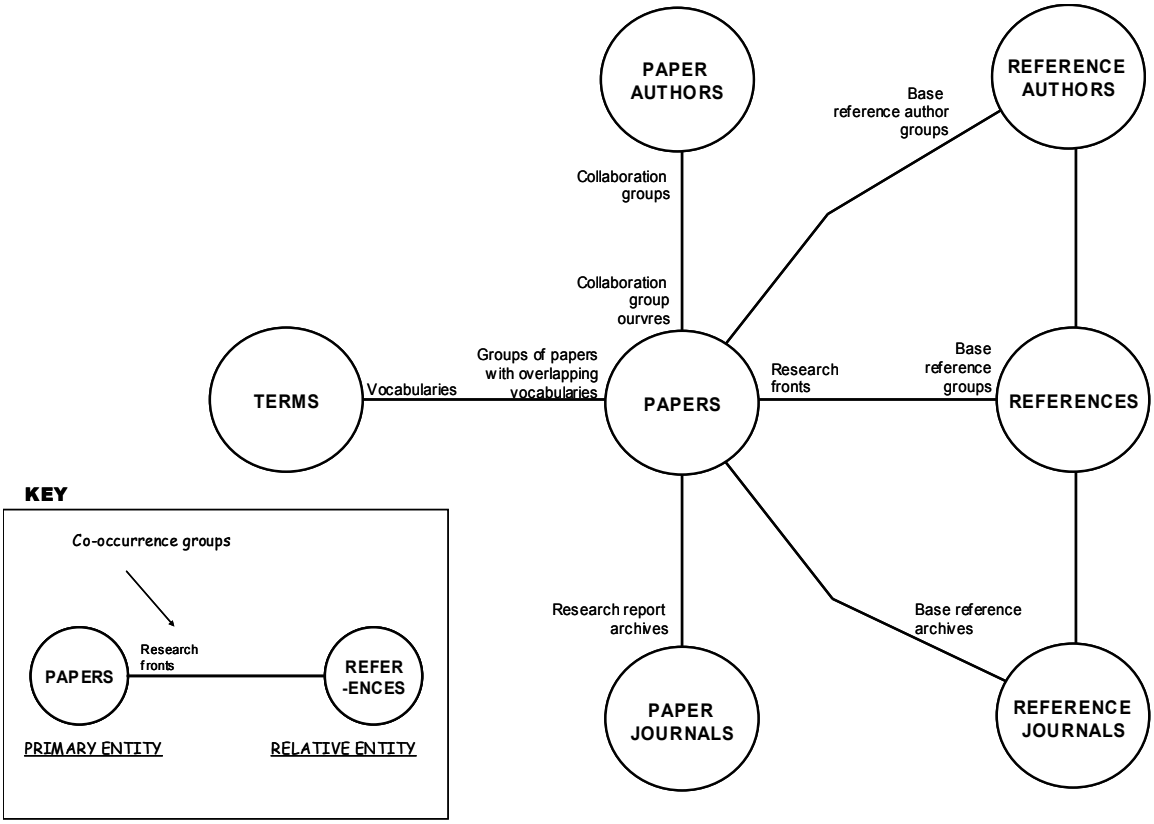


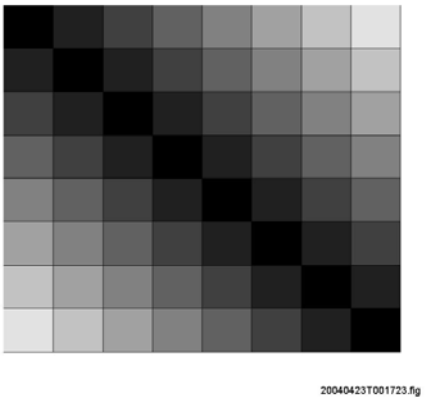
Figure 30. An entity relationship diagram showing symbolic representations for groups of entities formed by clustering on co-occurrence relations.

11.2 Seriation and matrix shading

Seriation is the method of ordering entities such that closely related entities are adjacent in the order and remotely related entities are far apart in the order. Seriation can be loosely considered a one dimensional multidimensional scaling problem (Kruskal & Wish, 1978). Seriation can be performed simultaneously on two entity-types related through an occurrence matrix by using a process known as *matrix shading*. The objective of matrix shading is to rearrange the rows and columns of a matrix such that it approximates, as much as possible, a *Robinson matrix*. In a Robinson matrix the magnitudes of the elements of the matrix decrease monotonically as one moves away from the matrix diagonal in any direction (Robinson, 1951). Visualizing the matrix by darkening matrix elements in proportion to their magnitude, a Robinson matrix is dark along the diagonal and gets lighter as one moves away from the diagonal. See Figure 31.

9	8	7	6	5	4	3	2	1
8	9	8	7	6	5	4	3	2
7	8	9	8	7	6	5	4	3
6	7	8	9	8	7	6	5	4
5	6	7	8	9	8	7	6	5
4	5	6	7	8	9	8	7	6
3	4	5	6	7	8	9	8	7
2	3	4	5	6	7	8	9	8
1	2	3	4	5	6	7	8	9

(a)



(b)

Figure 31. Example of the structure of a Robinson matrix (a), and the corresponding matrix shading (b).

Matrix shading algorithms typically use iterative matrix permutations to minimize a penalty function that measures the moment of the matrix element weights around the diagonal of the matrix. Assuming an occurrence matrix $\mathbf{O}[x_1;x_2]$, one such penalty function for matrix shading is:

$$\Psi = \sum_{i=1}^{nx_1} \sum_{j=1}^{nx_2} o_{ij}[x_1; x_2] \cdot (nx_2 \cdot i - nx_1 \cdot j)^2 \tag{102}$$

Figure 32 (a) shows an example of matrix shading of a paper to paper author matrix, $\mathbf{O}[p;ap]$, for a collection of papers covering the topic of SARS with 475 papers and 1901 paper authors. Papers authors with only one paper were eliminated, leaving a 475 paper by 443 binary paper author matrix. Matrix shading was performed using a greedy algorithm that iteratively performs alternating row and column permutations to minimize the penalty function of Equation (102). Dots in the figure correspond to ones in the matrix. Note that shading has concentrated the ones along the diagonal of the matrix, approximating a

Robinson matrix. Adjacent papers in this matrix are related because they share common paper authors. Adjacent paper authors in this matrix are related because they tend to be coauthors on the same papers.

Figure 32 (b) shows an example of matrix shading of a paper to reference matrix **O**[*p*;*r*], for a collection of SARS papers. There are 396 references in this collection. Reference authors appearing only once in the collection were eliminated, leaving a 475 paper by 584 reference binary matrix. Note in the figure how matrix shading tends to move the most highly cited references to the center columns of the matrix. These references appear as heavy vertical streaks in the shaded matrix. There is a great deal of overlap in references in adjacent papers, but the diagonal structure is definitely evident. In this case matrix adjacent papers in this matrix are related because they tend to cite the same references. Adjacent references in the shaded matrix are related because they tend to be cited by the same papers.

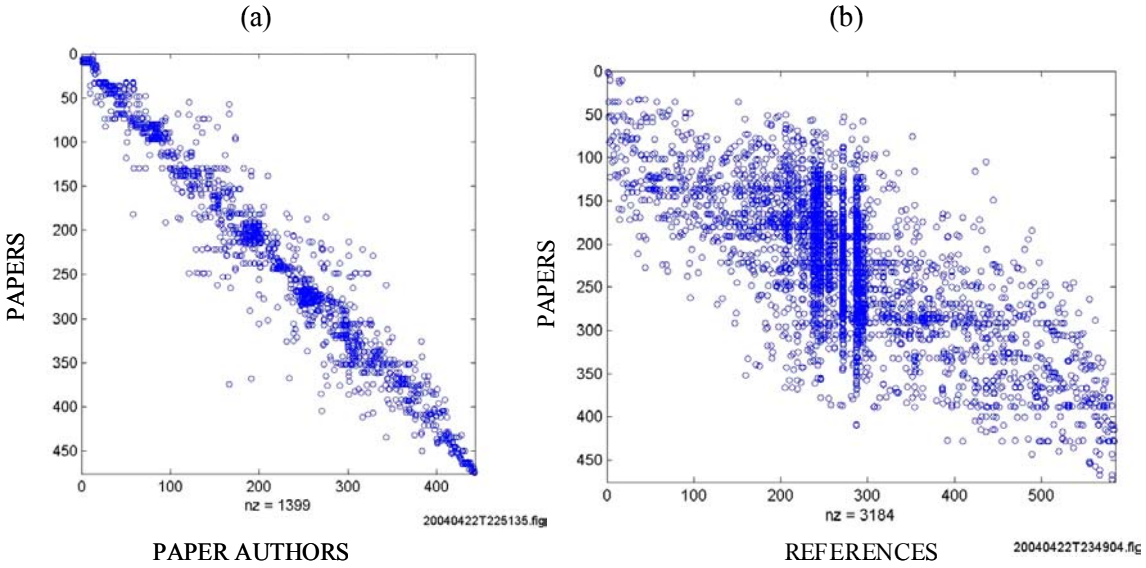


Figure 32. Results of matrix shading of occurrence matrices in a collection of papers on the topic of SARS research. Left in (a) shows a paper to paper author matrix. Right in (b) shows a paper to reference matrix.

Many methods exist for matrix shading. Packer (1989), for example, used simulated annealing and the penalty function of Equation (102) to do matrix shading on a paper author to reference author matrix, **O**[*ap*;*ar*], from a collection of papers on the topic of neuroscience. Several other matrix shading algorithms can be found in the literature (Brower & Kile, 1988; Lenstra, 1974).

11.3 Hierarchical agglomerative clustering

Agglomerative clustering gathers groups of entities by iteratively fusing clusters of entities that have the greatest similarity according to some *linkage function*. The method starts by assuming that each entity is a cluster with a single member. Given *N* entities in the collection, there are *N*-1 successive fusings, where

the members of the two most similar groups are combined, and the similarity of the newly fused group to all remaining groups is recalculated using the linkage function. Examples of commonly used linkage functions are single linkage, complete linkage, average linkage, and incremental sum-of-squares (Ward's method) linkage. Definitions of these linkage functions can be found in Gordon (1999). Careful selection of the linkage function is necessary in order to avoid *chaining artifacts*, where a large percentage of the entities are clustered into a single group (Gordon, 1999). The order of the successive fusings in agglomerative clustering can be used to build a clustering tree, or *dendrogram*. The tree can be truncated at the desired number of clusters.

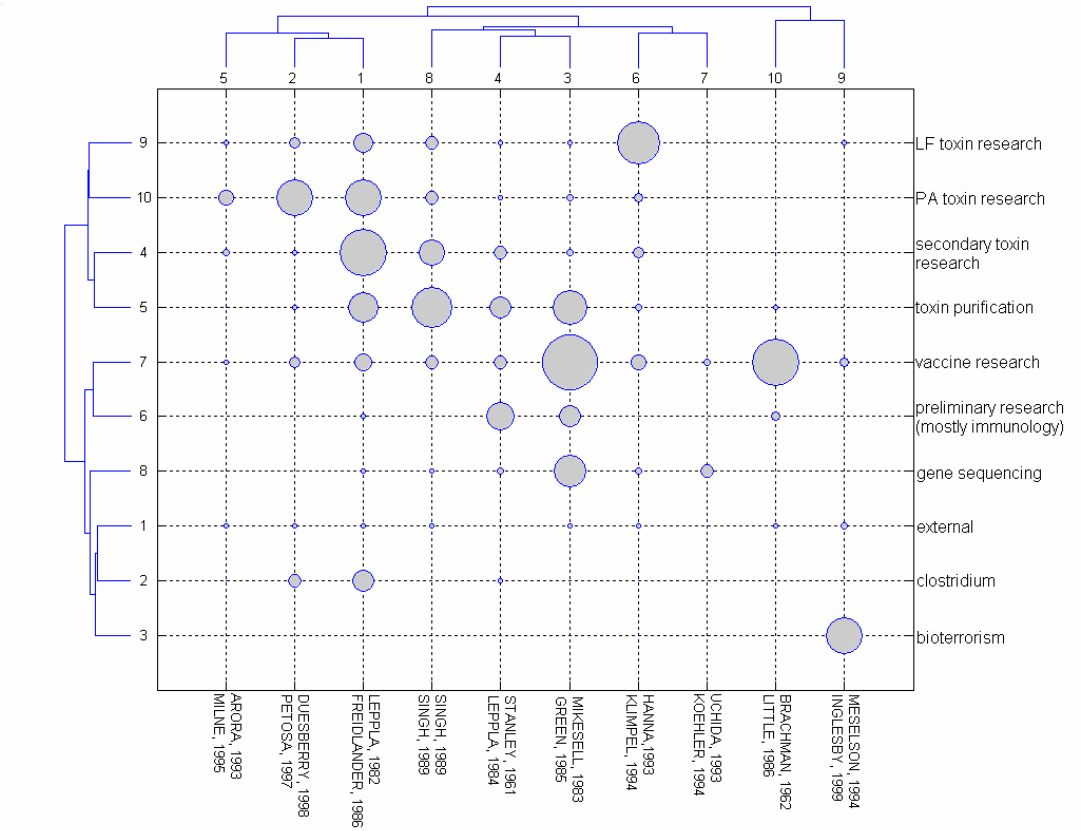


Figure 33. A crossmap of research fronts (groups of papers) to reference author groups for a collection of papers on the topic of molecular imprinting. The crossmap is a visualization of the paper to reference author matrix after agglomerative clustering and seriation of papers and reference authors. Note that the resulting matrix approximates a Robinson matrix.

The dendrogram also provides an ordering of the clusters. For any dendrogram with N leaves, there are 2^{N-2} possible orderings of the dendrogram. Selecting the ordering from among these possibilities poses another seriation problem. Seriation of dendrograms is a method often applied to genome data, for example Bar-Joseph, Gifford et al (2001) pose dendrogram seriation as a traveling salesman problem and use a search algorithm to find the optimal ordering. Morris, Asnake, and Yen (2003) use simulated annealing with a similarity times distance penalty function for dendrogram seriation. Agglomerative clustering and

dendrogram seriation can be applied to both primary and relative entities in an occurrence matrix. This method can be used for matrix shading and is the method used to produce crossmap visualizations (Morris & Yen, 2004) of occurrence matrices, which will be explained in Chapter 12.3. The net result of clustering entities is a shading of the occurrence matrix. For example, Figure 33 shows a research front (groups of papers by topic) to reference author group matrix, **O**[*gp*;*gar*] after clustering papers by bibliographic coupling and clustering reference authors by co-citation in papers. Here the circles on the diagram are proportional to magnitude of the elements of the matrix. Note the diagonal structure of the matrix after clustering.

11.4 Vector-based c-means clustering

Vector based clustering, such as c-means clustering, does not produce a dendrogram, as does agglomerative hierarchical clustering. In the c-means algorithm, the user selects the number of clusters *N*. The method starts by randomly selecting *N* points in the feature space. These are the cluster *centers.* At each iteration, two steps are performed: 1) the entities are assigned to the cluster whose center is closest to its feature vector in the feature space. and 2) each center is moved to the vector mean of the feature vectors of the entities assigned to its cluster. This iteration process proceeds until the cluster center positions converge to some final positions. The final result is: 1) a list of cluster memberships by entity, and 2) the cluster centers themselves, which can serve as prototypes that illustrate the typical pattern of features in each cluster. It is possible to use agglomerative clustering of the cluster centers to obtain a dendrogram if desired. Calculation times for c-means clustering tends to be shorter than agglomerative clustering. The method suffers less from chaining artifacts than agglomerative clustering as well. Chaining artifacts produce highly skewed distributions of cluster sizes in hierarchical clustering when entities do not fall into well-defined clusters in the feature space (Gordon, 1999). A fuzzy c-means algorithm (Bezdek, 1981) can be used that models fractional membership of entities in clusters, thus modeling overlap in cluster membership.

Similar to agglomerative clustering, c-means clustering can be considered as a rearrangement of the rows of the occurrence matrix. Using the cluster assignment produced by clustering and permuting the matrix so that rows from entities in the same cluster are adjacent to each other, the method produces a simple matrix shading that is much easier and faster to compute than using simpler matrix shading techniques discussed in Chapter 11.2.

Consider an example that uses the factor matrix from the well known study of information science authors by White and McCain (1998). The factor matrix is derived from factor analysis, a statistical technique, similar to latent semantic analysis, that expresses occurrences in terms of a reduced set of dummy entities. The technique is very similar to latent semantic analysis that is described in Chapter 13.8. Figure 34 (a) shows a diagram of the factor to reference author matrix, **O**[*f*;*ar*] from White and McCain (1998), where

dots in the diagram show non-zero elements in the matrix. Using agglomerative clustering, the factors were clustered based on the Euclidean vector distance between rows, while the reference authors were clustered based on the Euclidean vector distance between columns. After seriation of both the resulting dendrograms using simulated annealing (Morris, Asnake et al., 2003), the resulting entity orders were used to permute the factor to reference author matrix to yield the matrix in Figure 34 (b). Note the resulting structure in the permuted matrix, which approximates a Robinson matrix after a trivial reversal of column order. This illustrates that clustering of entities can be viewed as a method that produces matrix shading of occurrence matrices. Figure 35 shows a permuted matrix of 34 (b) with clustering dendrograms, factor labels and reference author labels attached. It is evident, from the dendrograms and the structure of overlap of reference authors across the factors, that factors and reference authors generally fall into two fields, information retrieval on the bottom and left of the matrix, and bibliometrics/citation theory on the top and right of the matrix, broad classifications discussed by White and McCain (1998).

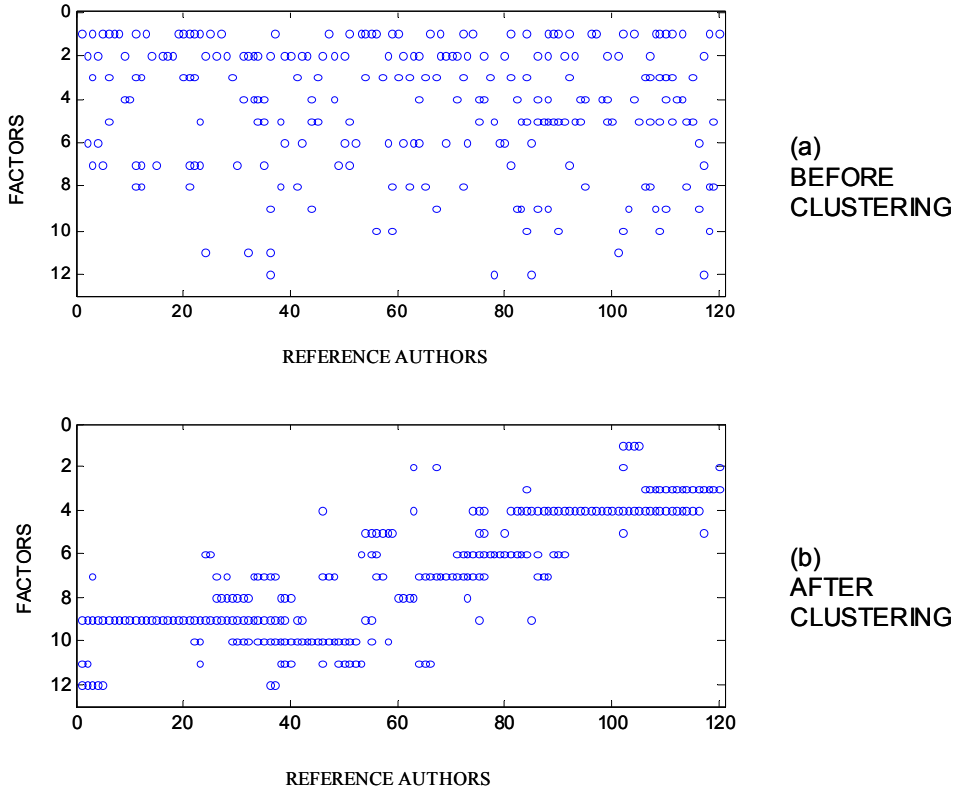


Figure 34. Factor to reference author matrix from analysis of Information Science authors by White and McCain (see text). Top (a) shows the original matrix. Bottom (b) shows the matrix after agglomerative clustering and seriation.

All the techniques discussed in this chapter: matrix shading, seriation, hierarchical clustering and c-means clustering can be seen as methods to rearrange the rows and columns of an occurrence matrix to approximate a Robinson matrix, thereby exposing the similarity structure of the entities of both the primary and the relative entity-types under consideration. Chapter 6, which follows, will discuss how this matrix shading can be exploited to visualize the structure of links among entities in a collection of papers.

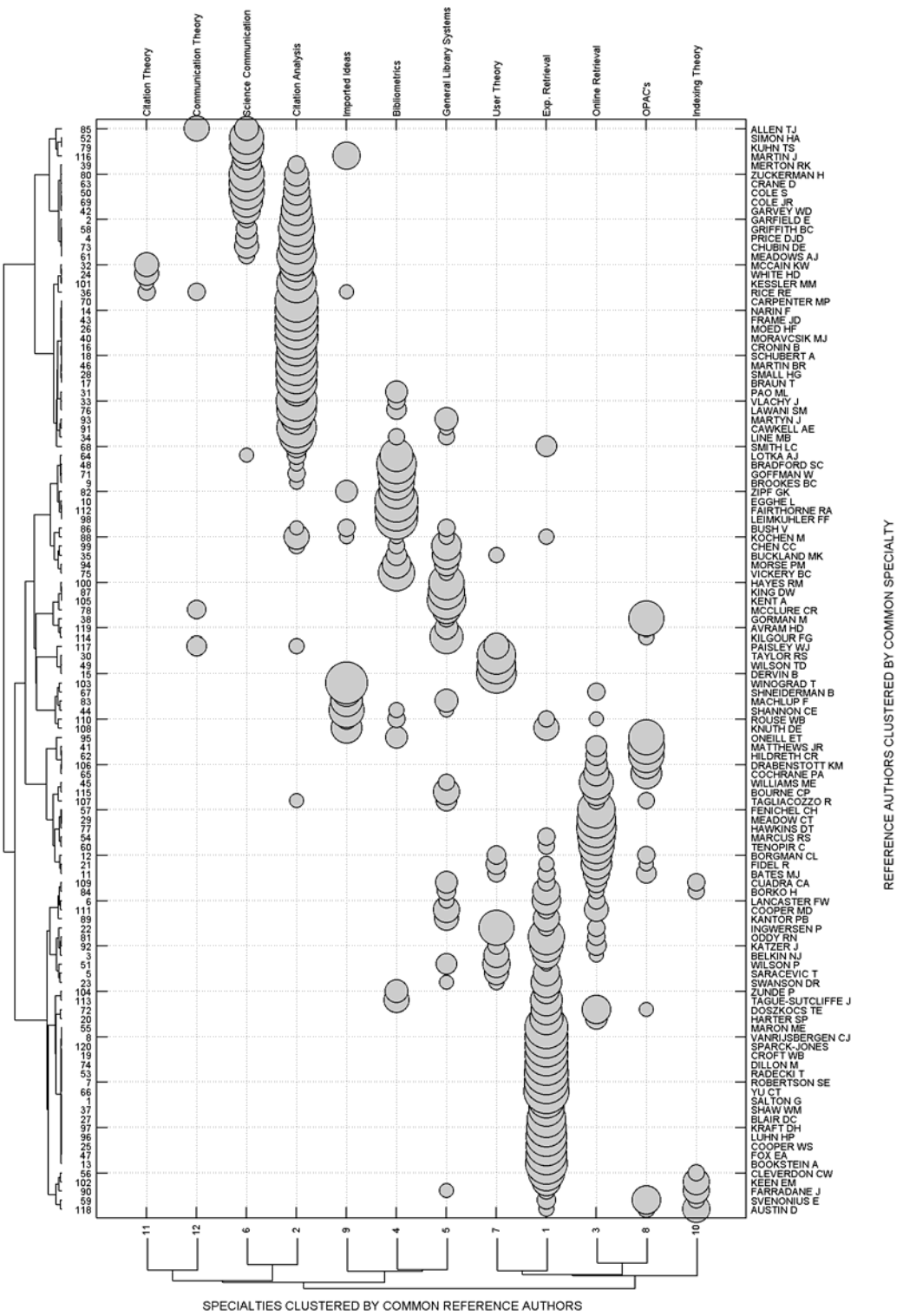


Figure 35. Factor matrix of Figure 34 (b) with clustering dendrograms and group labels.

12. VISUALIZATION OF OCCURRENCE MATRICES

Three types of visualizations are found useful for mapping relations within a collection of papers: 1) timelines, 2) usage plots, and 3) crossmaps. Each of these visualizations is a simple visual representation of occurrence matrices within a collection of papers. Matrices are most readily visualized as *bubble plot*s, in which the elements of the matrix are plotted as bubbles whose size is proportional to the magnitude of the matrix element (or some other variable) and whose position on the plot is determined by its row and column position in the matrix.

12.1 Timelines

Timelines are maps of individual entities plotted by time. Entities are mapped as dots on the plot, placed in horizontal tracks by group on the y axis and plotted by time on the x axis. Timelines are typically drawn for groups of papers in an effort to visualize research trends, emergence of new research fronts, and obsolescence of old research fronts. Assume that a membership matrix **O**[*gp;p*] is constructed using some clustering technique to group the papers (denoted *gp*) into research fronts (groups of papers that cover a similar research topic.) For example, hierarchical agglomerative clustering can be used to cluster papers into research fronts using bibliographic coupling (Morris, Yen, Wu, & Asnake, 2003). Using the paper publication year, *yp*, as an entity-type, the occurrence matrix **O**[*p;y*p] lists the publication year of the papers in the collection. The timeline matrix can be formed by multiplying out the matrices:

$$\mathbf{O}[gp;py] = \mathbf{G}[gp;p] \cdot \mathbf{O}[p;yp] \tag{103}$$

A similar matrix can be formed for publication month, *mp*, if such data is available for papers:

$$\mathbf{O}[gp;mp] = \mathbf{G}[gp;p] \cdot \mathbf{O}[p;mp] \tag{104}$$

A typical timeline is shown in Figure 36, from a collection of papers on the topic of anthrax research published from 1945 to 2002. In this case the papers have been clustered into 35 groups using similarities based on bibliographic coupling. Papers were excluded that did not have a minimum of 5 bibliographic coupling counts with at least one other paper in the collection. The papers have been plotted according to Equation (104) but the plot has some added information. The size of the circles representing each paper is proportional to the number of times that the paper has been cited. Circles are shaded to be proportional to

the number of times that a paper has been cited from papers in the final 12 months of the collection. Note that with this visualization that it is possible to pick out important papers in the collection based on the number of times each paper has been cited. Furthermore, it is easy to pick out current high-interest, highly cited papers by their dark shading on the plot. In this visualization the research front appears as horizontally arrayed circles with corresponding topic labels on the right of the plot. It is easy to see the emergence of individual research fronts on the timeline. Obsolescence of research fronts and superceding of old research fronts by newer ones is also easily distinguished on this visualization.

Timelines of groups of base references can be used to find sources of loan knowledge for emerging specialties. Assuming that groups of references are formed using some clustering technique to yield a membership matrix **G**[*gr*;*r*], and that a reference to reference year matrix, **O**[*r*;*yr*], is constructed, the reference group to reference year matrix can be calculated from :

$$\mathbf{O}[gr;yr] = \mathbf{G}[gr;r] \cdot \mathbf{O}[r;yr] \tag{105}$$

This is a matrix of groups of references as the rows, and reference year as columns. For any reference year column, the elements of the matrix are the count of the number of references in a group that have that reference year.

In emerging specialties, references that are used to symbolize loan knowledge from other fields usually predate the date of emergence of the field. Figure 37 shows a base reference timeline for a collection of papers on the topic of complex networks. Here, the references have been clustered into base reference groups using co-citation. Reference years are on the y-axis and the base references are arrayed on the x-axis. Putting the base references on the x-axis allows the base references to be aligned with crossmap visualizations where base references are also shown on the x axis.

Note that in this collection the principle discovery paper for the specialty is a 1998 paper by Watts and Strogatz, whose corresponding reference is the large circle marked with crosshairs on the figure. References that pre-date the Watts and Strogatz discovery reference should represent loan knowledge drawn from other specialties. As shown in the figure, there are four base references that pre-date the Watts and Strogratz reference. On the left are references corresponding to a 1967 paper by Milgrams on social 'small world phenomena,' and a second reference to a 1994 paper by Wasserman on social network analysis. These two references represent loan knowledge drawn from social network theory. Two references on the right, one by Erdos, another by Bollobas, represent loan knowledge drawn from random graph theory.

12.2 Usage plots

Usage plots are similar to timelines, but in this case, instead of publication year, the number of associations between two groups of entities is plotted as a function of time. Usage plots of references in a collection can be used to map the accretion of knowledge in a specialty, as shown by the appearance of exemplar references in the specialty. This type of plot can be calculated from the occurrence matrices:

$$\mathbf{O}[gr;yp] = \mathbf{G}[gr,r] \cdot \mathbf{O}[r,p] \cdot \mathbf{O}[p,yp] \tag{106}$$

This matrix has groups of references as rows and paper publication years as columns. For any particular publication year, the matrix elements count the number of citations to references within a reference group for that year. This visualization technique shows the emergence and obsolescence of groups of exemplar references that accompanies discoveries and paradigm shifts. Figure 38 shows a usage plot of references from a collection of papers on the topic of angiogenesis. In this plot it is possible to identify the growth of specific references as exemplar references in the collection. Exemplar references appear as vertical tracks of large circles in the plot. The time of emergence of concepts in the specialty can be seen as the starting time of the tracks of the well-cited exemplar references. Paradigm shifts and obsolescence are seen as sudden declines of citations to formerly well-cited references, although the usage plot shown in the figure does not contain good examples of such sudden citation decline.

Usage plots of reference authors in the collection can be used to map *schools of thought*, that is broad conceptual frameworks typically associated with specific groups of authors. Assuming groups of reference authors, *gar*, formed by clustering on co-citation of authors in papers, the reference author group to paper year matrix can be calculated by matrix arithmetic from:

$$\mathbf{O}[gar;yp] = \mathbf{G}[gar,ar] \cdot \mathbf{O}[ar,r] \cdot \mathbf{O}[r,p] \cdot \mathbf{O}[p,yp] \tag{107}$$

This matrix, with reference authors as rows and publication year as columns, shows the number of citations received by the reference authors in each reference author group by year. Figure 39 shows a usage plot of reference authors from a collection of papers on the topic of angiogenesis. Similar to a reference usage plot, in this plot it is possible to identify the growth of specific reference authors as symbols of schools of thought in the collection as vertical tracks representing highly cited reference authors. Also, similar to reference usage plots, paradigm shifts and obsolescence of schools of thought can be tracked from temporal changes in the number of citations received by base references authors.

12.3 Crossmaps

Crossmaps are visualizations of occurrence matrices between pairs of entity-types (Morris & Yen, 2004). Crossmaps are very useful for visualizing overlap in relations among groups of entities. Entity groups for both entity-types are formed by agglomerative hierarchical clustering. After clustering, dendrogram seriation is performed to place related leaves adjacent to each other. The rows and columns of the group co-occurrence matrix are rearranged to match the two dendrograms. The resulting matrix tends to approximate a Robinson matrix. The matrix is plotted as a bubble plot with the dendrograms appropriately placed at the left and top, and group labels placed on the right and bottom. In this type of plot, the overlap of relations across groups of different entities tends to show as clumps of bubbles on the plot.

One type of useful crossmap can be formed from groups of papers and groups of references. Assume membership matrix **G**[*gp;p*], which lists of groups of papers (research fronts) clustered using bibliographic coupling, and also assume membership matrix **G**[*gr*;*r*], which lists groups of references *gr*, as exemplar reference groups clustered using co-citation, The occurrence matrix for plotting can be calculated from:

$$\mathbf{O}[gp;gr] = \mathbf{G}[gp,p] \cdot \mathbf{O}[p,r] \cdot \mathbf{G}[r,gr] \tag{108}$$

This type of crossmap shows the relation between research fronts (groups of paper covering the same topic) and exemplar reference groups (groups of references that are often cited together in papers.) As such, it is a visualization of the relation of base knowledge to research in the collection. Figure 40 shows a crossmap of research fronts to base references for a collection of papers covering the topic of angiogenesis. In this crossmap groups of large circles correspond to groups of papers that are closely related to groups of exemplar references.

A second type of useful crossmap is derived from the occurrence matrix of groups of papers and groups of reference authors. Papers are clustered by bibliographic coupling into research fronts to yield membership matrix **G**[*gp;p*], while reference authors are clustered by co-citation counts to yield membership matrix **G**[*gra*;*ra*]. Research fronts to reference author groups can be calculated from:

$$\mathbf{O}[gp;gar] = \mathbf{G}[gp,p] \cdot \mathbf{O}[p,r] \cdot \mathbf{O}[r,ar] \cdot \mathbf{G}[ar,gar] \tag{109}$$

Figure 41 shows an example of a crossmap of research fronts to base reference authors. The interpretation of this type of crossmap is similar to that of the research front to base reference group crossmap shown in Figure 40. The research fronts represent research topics while the base reference author groups represent schools of thought.

A third type of useful crossmap is derived from the occurrence matrix of groups of papers and groups of paper authors. Assume **G**[*gp*;*p*], the paper membership matrix discussed above, and also assume **G**[*gap*;*ap*], a paper author membership matrix from groups of paper authors clustered using co-authorship counts. Research fronts to paper author groups can be calculated from:

$$\mathbf{O}[gp;gap] = \mathbf{G}[gp,p] \cdot \mathbf{O}[p,ap] \cdot \mathbf{O}[ap,gap] \tag{110}$$

Figure 42 shows a crossmap of reference fronts to paper authors for a collection of papers covering the topic of angiogenesis. In this map collaboration groups appear as leaves on sub-branches of the dendrogram that is placed along the top of the plot. Groups of horizontal adjacent circles on the map correspond to collaborating groups of authors publishing a series of papers in a research front.

The visualization techniques described in this chapter are effective in showing the static structure of links as well as dynamic changes in link structure in a collection of papers. Timelines of papers can be used to show emergence and obsolescence of ideas, important papers and current hot concepts. Timelines of references can show references corresponding to ideas borrowed from other fields. Usage plots allow the easy visualization of growth of key concepts and schools of thought in a specialty. Crossmaps show links among unlike entities, and particularly show the relation of research fronts to schools of thought, groups of exemplar references and author collaboration groups.

All of these visualizations are adaptations of simple bubble diagrams of occurrence matrices. As shown by the equations presented in this chapter, these matrices are easily calculated from direct occurrence matrices using matrix arithmetic.

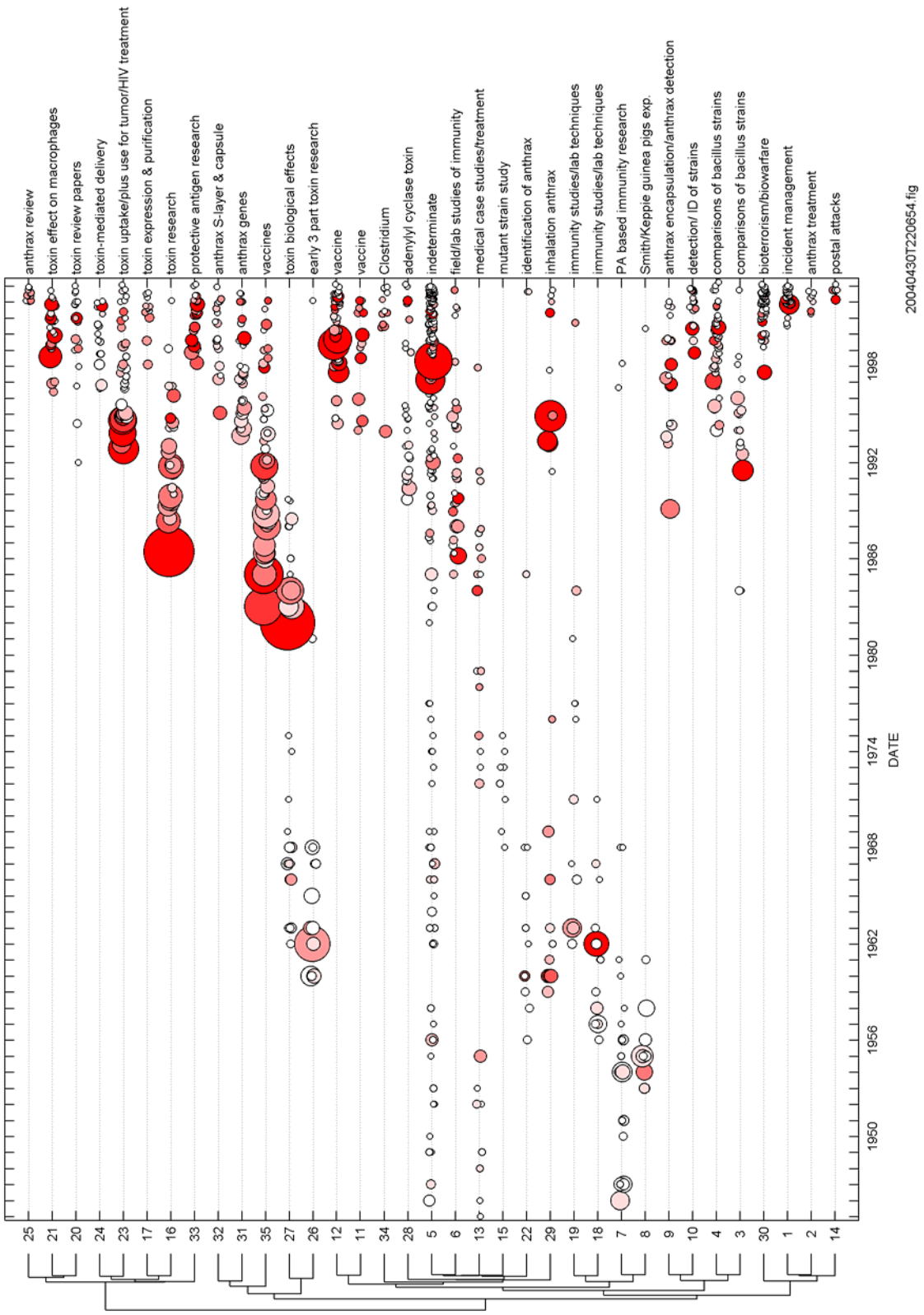


Figure 36. Research front timeline of a collection of papers on the topic of anthrax research over a 60 year period.

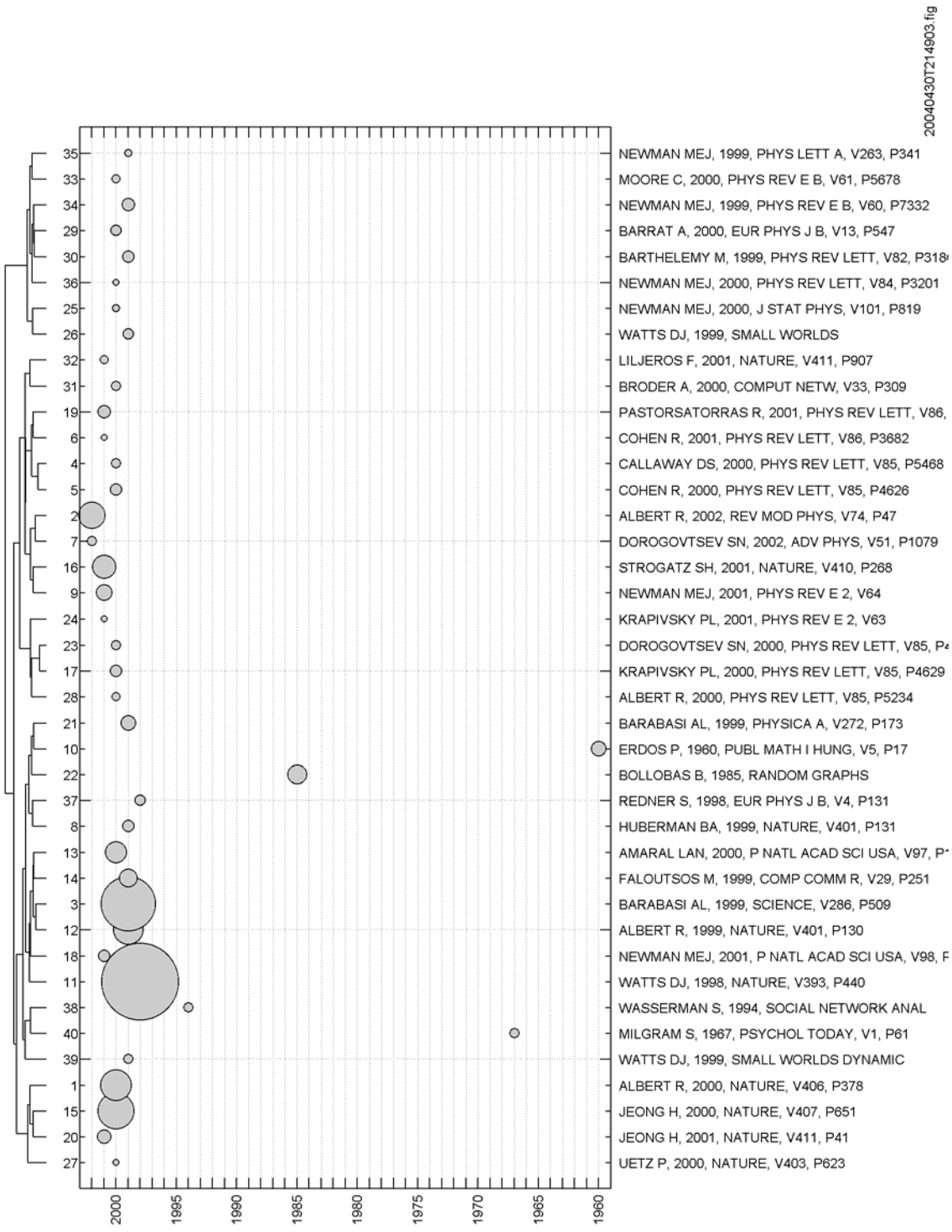


Figure 37. Timeline of base references for a collection of papers on the subject of complex networks.

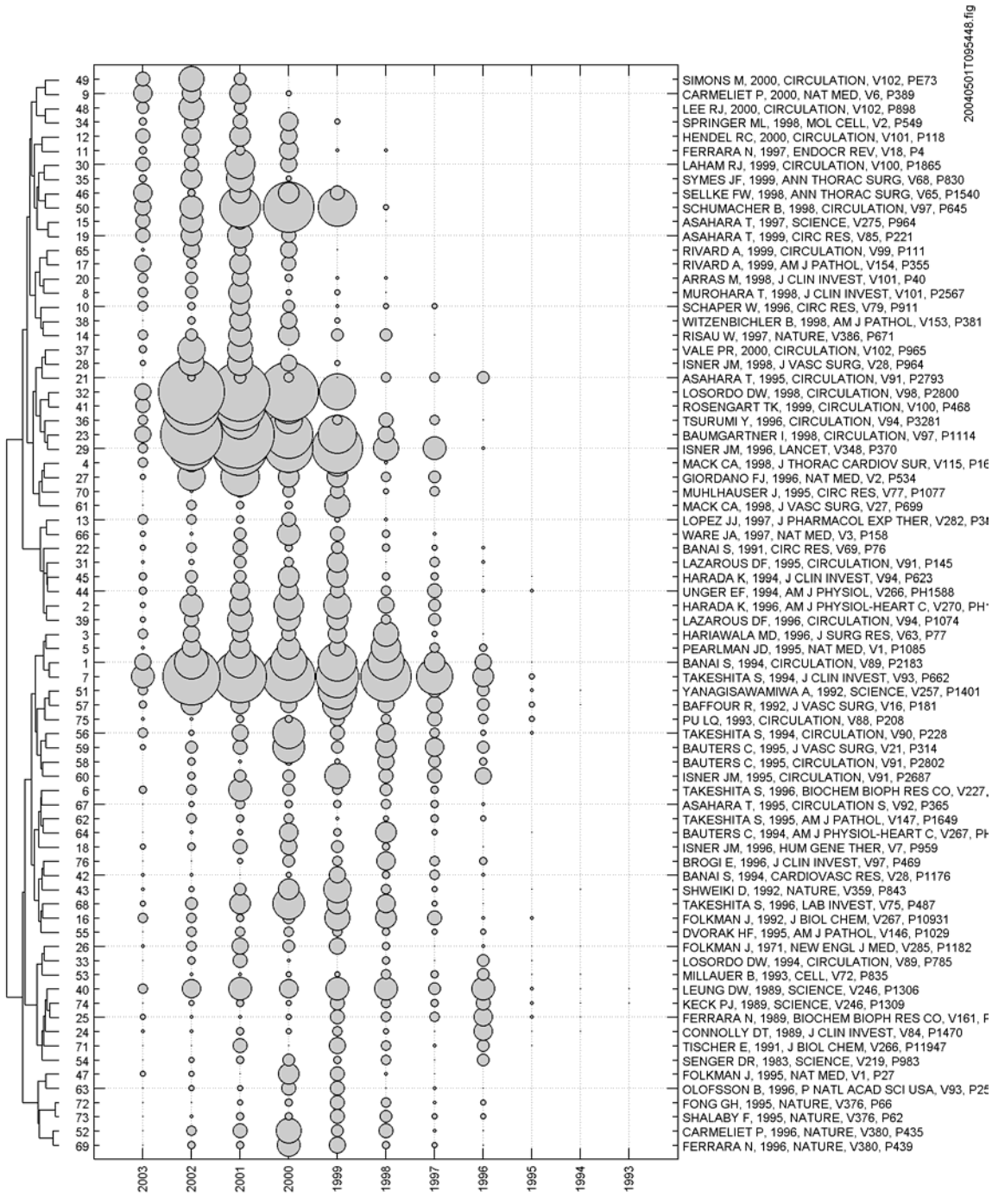


Figure 38. Reference usage plot from a collection of papers on the subject of angiogenesis.

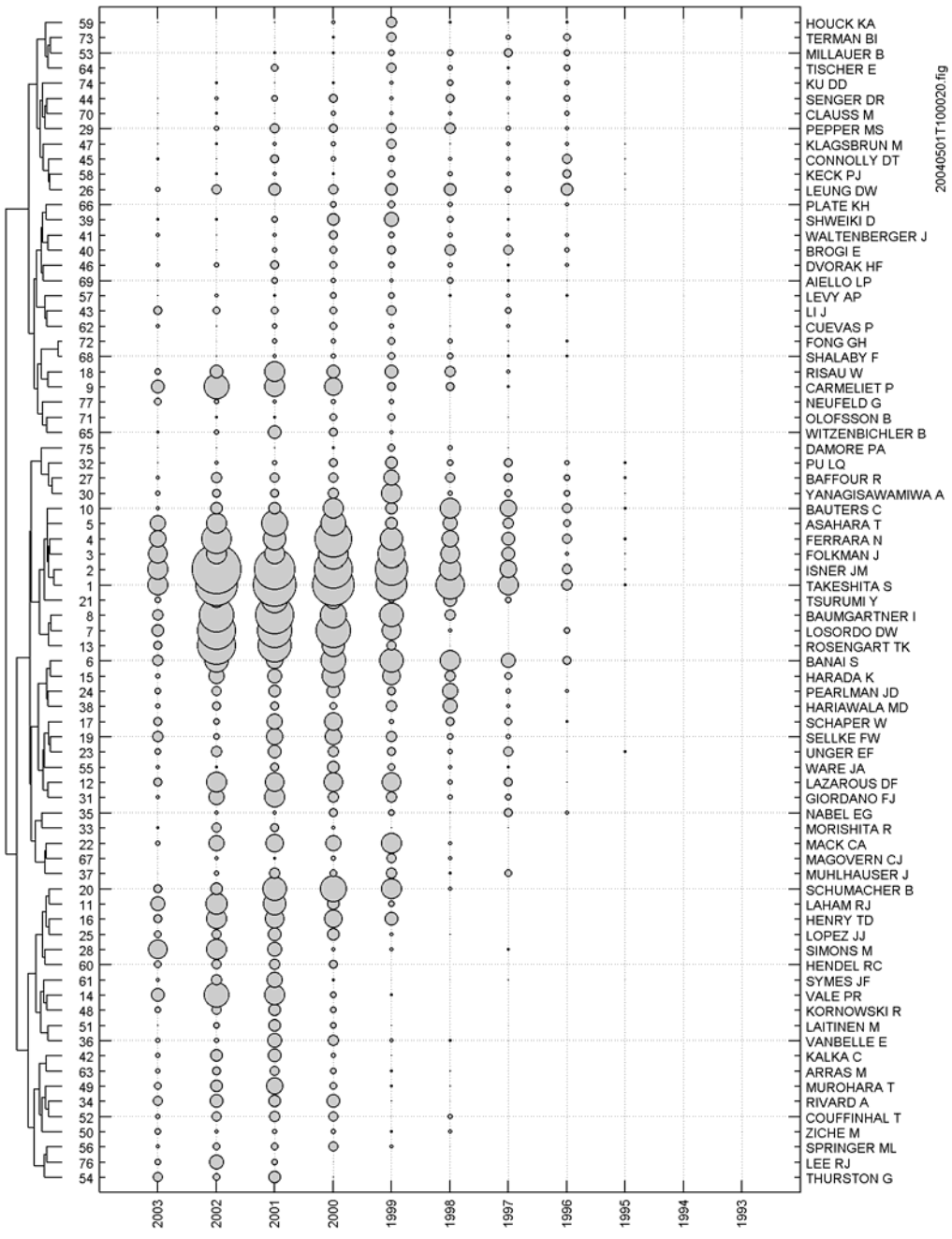


Figure 39. Reference author usage plot from a collection of papers on the subject of angiogenesis.

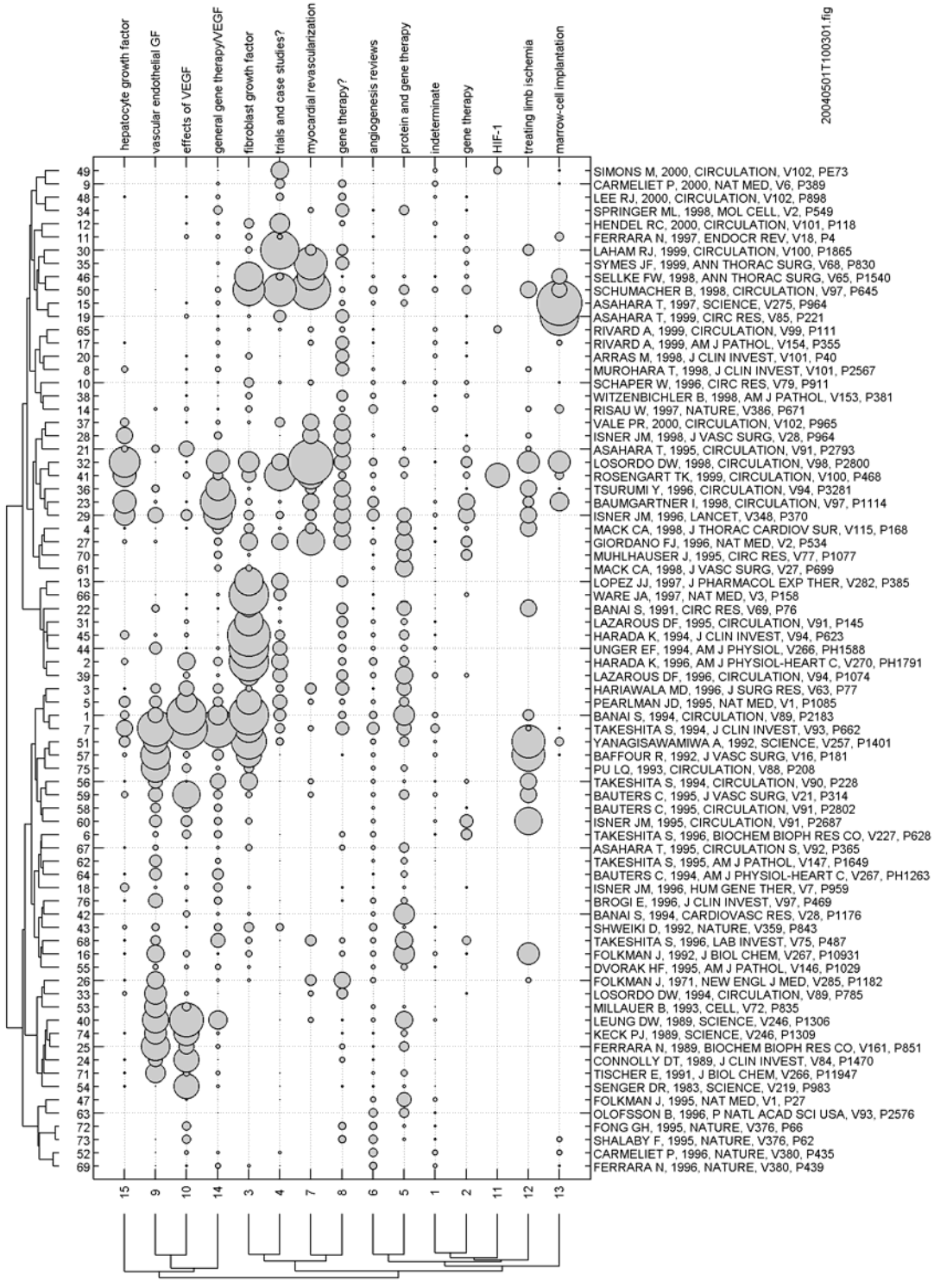


Figure 40. Research front to base reference crossplot from a collection of papers on the topic of angiogenesis.

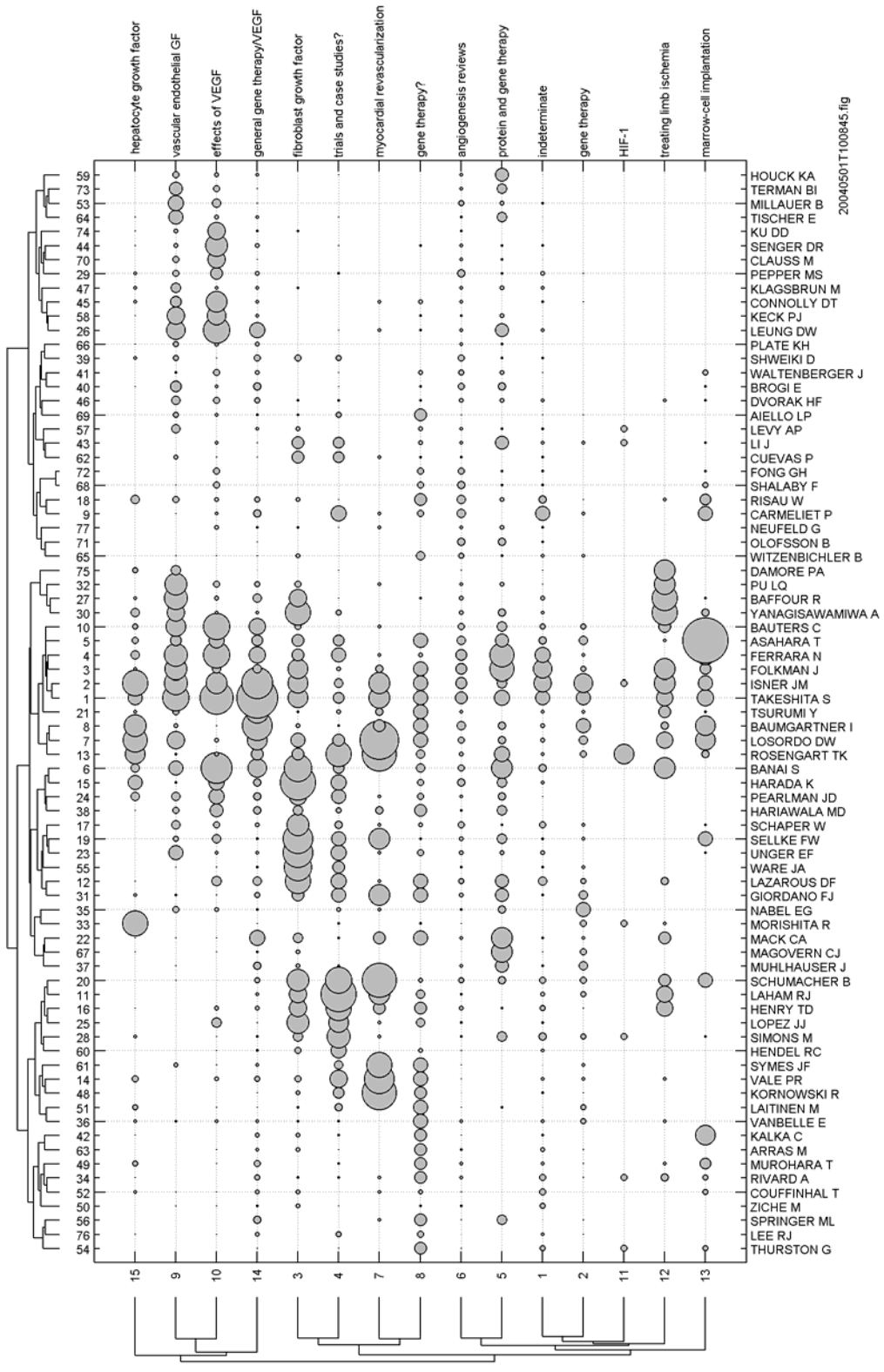


Figure 41. Research front to base reference author crossmap for a collection of papers on the subject of angiogenesis.

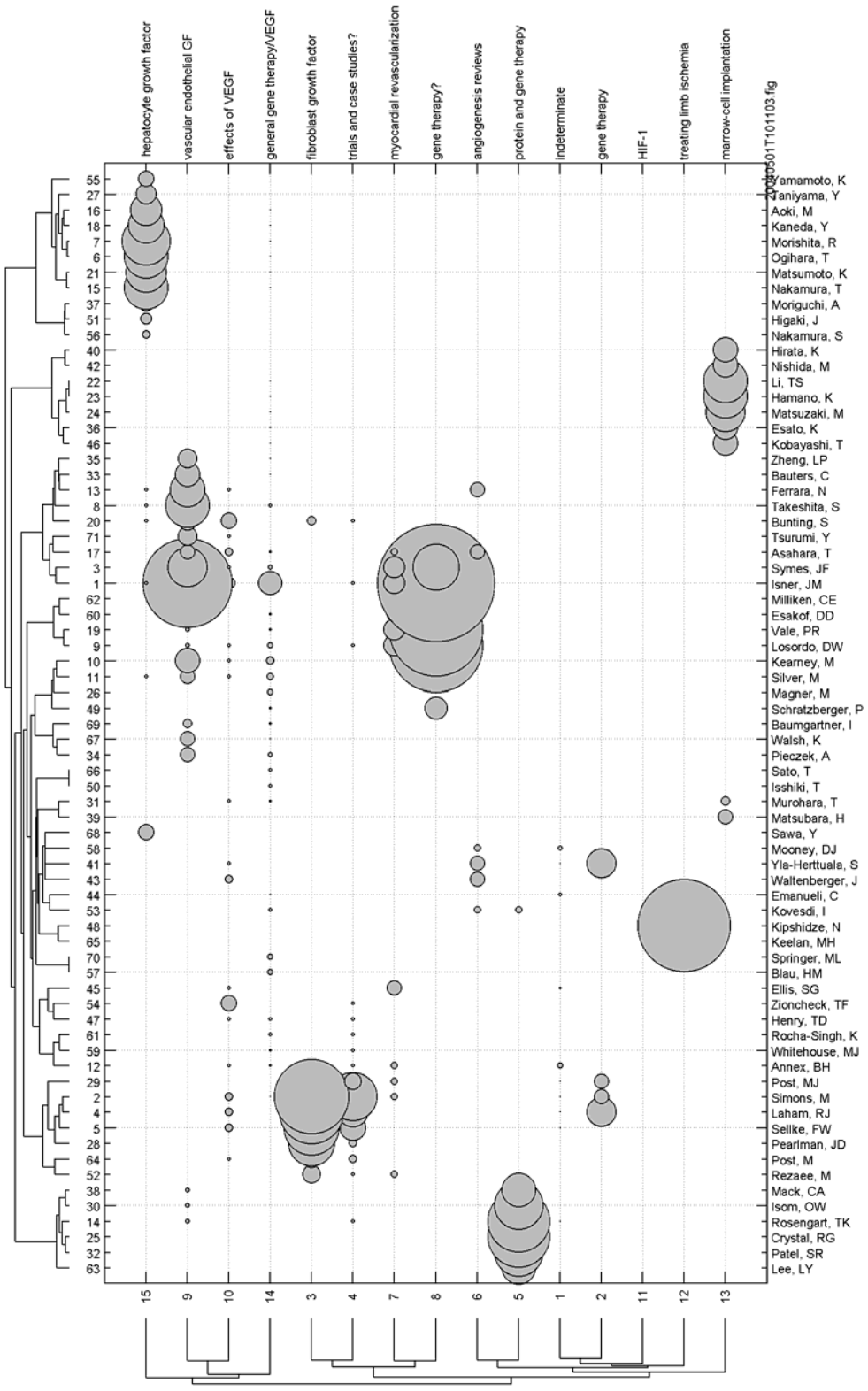


Figure 42. Research front to paper author crossmap for a collection of papers on the subject of angiogenesis.

13. EXISTING ANALYSIS TECHNIQUES

13.1 Introduction

This chapter will review many techniques commonly used in bibliometric analysis and present them in the context of the mathematical treatment proposed in this report. All of these methods can be expressed in terms of the formation and manipulation of occurrence matrices and co-occurrence matrices. The following bibliometric analysis techniques will be reviewed:

- **Co-occurrence clustering techniques**
  - **Co-citation analysis**. This is a technique for finding groups of related references that are cited together in papers. The technique is used for finding groups of base references.
  - **Bibliograhic coupling analysis**. A clustering and searching method for finding groups of papers that tend to cite the same references. This technique groups papers by topic.
  - **Author co-citation analysis**. A method of mapping a field by finding groups of reference authors that tend to be cited in papers together.
  - **Journal co-citation analysis**. A method of finding groups of related journals that tend to be cited together. This technique is used to find base reference archives.
  - **Braam-Moed-vanRaan (BMV) co-citation/co-word analysis**. A method to relate groups of references that are cited in papers together to groups of terms.
- **Latent variable and modal analysis techniques**
  - **Latent semantic indexing**. A dummy entity technique where term occurrences are expressed in relation to dummy concepts.
  - **Hubs and authorities analysis**. A search method, originally developed for searching on the World Wide Web, For finding groups of authorities (references that tend to be cited together by hubs) and authorities (papers that tend to cite the same authorities.)
- **Feature vectors**
  - **Author identities and author images**. A technique for characterizing individual reference authors and paper authors. This technique finds "schools of thought."
- **Network pruning techniques**

- **Pathfinder analysis**. A method of link pruning used to visualize network structure.

The discussion to follow will center on showing that these techniques can be expressed in terms of occurrence and co-occurrence matrices and showing that the algorithms associated with each of these methods can be expressed simply in terms of the mathematical treatment proposed here.

13.2 Co-citation analysis

Co-citation analysis is the most widely used method to map knowledge in a collection of papers. This is done by clustering or mapping references in a collection of papers based on co-citation counts. Most common applications use a citation threshold to select references that have been heavily cited. Using a reduced co-citation matrix, applications such as multidimensional scaling (MDS), pathfinder maps, and agglomerative hierarchical clustering are used to find the structure of relations among references. Clusters of references mapped this way constitute groups of *exemplar references* or *base references*, references that are used as knowledge symbols by authors publishing papers within a specialty. If such highly cited references are considered exemplars, then these groups of references that are highly cited together can be thought of as exemplars that are part of common Kuhnian paradigms. (Morris, 2004; Small, 1978)

Using the mathematical treatment proposed here, co-citation clustering can be presented as:

$$\mathbf{G}[gr;r] = f(\mathbf{C}[r;p]) \tag{111}$$

where *gr* represents *base reference groups*, and $f(\mathbf{C}[r,p])$ is a clustering function such as hierarchical agglomerative clustering or c-means clustering. Small (1997), for example, describes a co-citation clustering technique based on hierarchical agglomerative clustering with single linkage where a maximum is enforced on the number of cluster members. This membership matrix can be used to relate base reference groups to other entities or entity groups in the collection of papers. See, for example, the co-citation/co-word technique discussed in Chapter 13.7.

13.3 Bibliographic coupling analysis

Clustering of papers using bibliographic coupling was introduced by Kessler (1963). As previously noted, bibliographic coupling occurs between pairs of papers and is the count of the number of references that are cited by both papers. It is assumed that pairs of papers that cite many common references cover similar research topics. From this assumption, it can be inferred that groups of papers that are clustered based on bibliographic coupling are papers that cover the same research topic. This is the assumption used by the ISI in their Web of Science product to find papers that are related to a paper selected by the user.

Bibliographic coupling is a static metric, as shown in Chapter 7.4, the bibliographic coupling count between a pair of papers never changes. Morris, Yen et al (2003) exploit this characteristic to use

bibliographic coupling to cluster papers into research fronts that are displayed in timelines to visualize temporal events in the literature of a specialty. These events include trends, discoveries, obsolescence, and topic fission. Research fronts formed through bibliographic can coupling be visualized in a crossmap format, that shows relations of research fronts to other entity groups within the collection of papers (Morris & Yen, 2004). For a better description of timelines and crossmaps, see Chapter 6.

13.4 Discussion of research fronts and exemplar reference groups

The concept of a research front is important when studying literatures. There is a need to classify papers by both topics and by currency. The term *research front* was originally defined by Price (1965). Price's model of literature was of papers citing papers, that is, a graph theoretic model as in Chapter 8, with no distinction between papers and references. Price defined a research front as the "growing epidermis" of papers and the papers in the immediate past that they cite. Price used a cited paper to citing paper matrix (an adjacency matrix **O**[*cp*;*p*] as defined in Chapter 8.) An example of such a matrix, taken from a collection of papers on the topic of angiogenesis is shown in Figure **43**. Note the dashed line drawn about 50 papers above the diagonal of the matrix. Assuming some paper *i* as the latest paper published, Price considered the research front at that instant to be paper *i* and the 50 papers or so that preceded it. On the adjacency matrix of Figure **43**, this defines the research front at paper *i* as the papers in the triangular area that trails it. The importance of the concept of the research front is that it attempts to identify what is current in the literature, in effect, "high-grading" the papers in the literature and providing a simple model of current and obsolete literature. This model of research fronts is difficult to apply practically, and has been superceded by other definitions

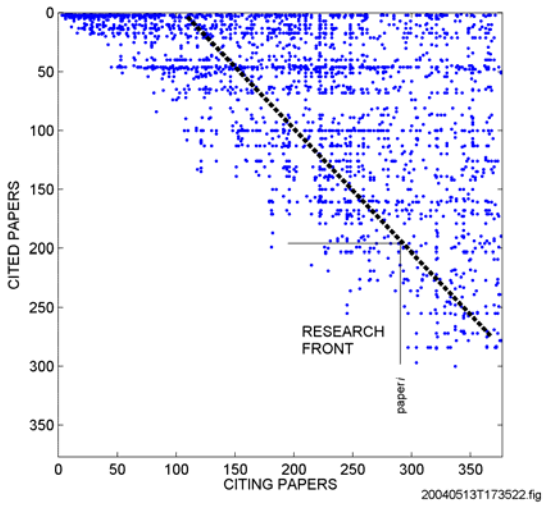


Figure 43. Example of a research front as defined by Price, on an paper adjacency matrix from a collection of papers on the topic of angiogenesis. The research front consists of the 50 papers immediately preceding the most current paper published.

Garfield (1994) defined research fronts as "co-citation clusters and the papers that cite them." This is a useful operational definition. Assume a collection of papers where papers are clustered into groups using bibliographic coupling, and references are clustered into groups using co-citation. These groups form the occurrence matrix **O**[*gp*;*gr*]. Figure **44** shows an example of this type of matrix, taken from a collection of papers on the topic of MEMS RF switches. References correspond to columns in this matrix and papers correspond to rows. A group of 4 clustered references is identified at the top of the figure. A dashed box is drawn around the columns for this group of references. Groups of papers that cite the references are noted by arrows on the right. By the Garfield definition, the group of 4 references and the papers that cite them constitute a research front. A problem with this definition is that it mixes references and papers. References are *concept symbols* (Small, 1978) or represent knowledge sources and can be ranked in importance by the number of citations they receive. Papers are simple research reports and cannot be ranked.

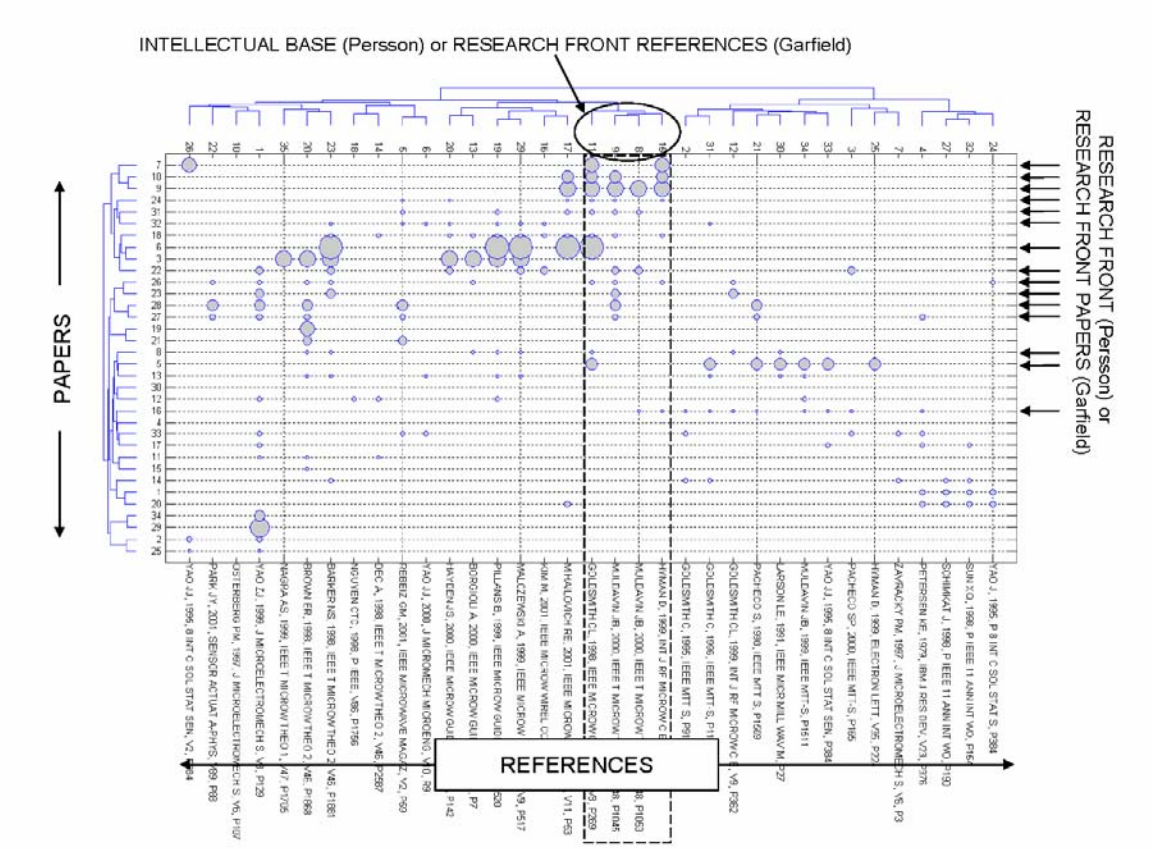


Figure 44. Crossmap of paper groups (clustered using bibliographic coupling) by reference groups (clustered using co-citation) from a collection of papers on the topic of MEMS RF switches. Under Garfield's definition a reference front is a group of references, such as the 4 references shown, and the papers citing them, such as the papers in groups identified using arrows on the right. Persson defines an intellectual base as a co-citation cluster of references (such as the column of 4 references highlighted) and a research front as the papers that cite the co-citation cluster, such as the citing papers in the paper groups noted on the right.

Persson (1994) separates the references from the papers when defining research fronts. Assuming a group of references clustered using co-citation, a research front is defined as the papers that cite that group, which is defined as the *intellectual base* of the research front. Examining Figure **44** again, considering the same group of 4 references as before, the column of 4 references represents the intellectual base of a research front, while the papers that cite those references, indicated by the arrows on the right, comprise the associated research front. This separation allows consideration of the research front's papers as simple reports covering a topic, while allowing separate consideration of the intellectual base references as knowledge sources and concept symbols. Note, however, that papers themselves are not grouped. There is the problem of establishing research front membership for emerging specialties and sub-topics. In this case the heavily cited references are not well established and the co-citation clusters are changing rapidly as exemplar references accrete in the specialty (Morris, 2004)

Morris, Yen, et al (2003) use a definition of research front similar to Persson's. In this case however, research fronts are defined as groups of papers clustered by bibliographic coupling. *Base documents* of research fronts are defined as the references that are cited by 40% or more of the papers in a research front. In a second paper, Morris and Yen (2003) introduced an entity-relationship model of collections of journal papers and proposed research fronts as papers clustered by bibliographic coupling and base reference groups as references grouped by co-citation, as shown in Figure **45**. Using bibliographic coupling to group papers into research fronts introduces temporal stability into research fronts because links between papers are static and do not cumulate (Morris, Yen et al., 2003) Clusters of papers found using bibliographic coupling should be more robust than when using Persson's definition because they are gathered using links among themselves rather than gathering papers that are indirectly related by citing co-citation clusters of references. There is a great deal of overlap of research fronts and the base reference groups they cite. However, the crossmapping technique introduced by Morris and Yen (2004), and explained in Chapter 12.3 is useful for visualizing these overlapping relations between research fronts and base reference groups, and adds considerable insight into the analysis of collections of journal papers.

Research fronts are unique in their importance to the analysis of a specialty through a collection of journal papers. In a broad sense, research fronts are more than just groups of papers clustered by topic. The authors of the papers in the research front tend to cite the same references and are therefore drawing on the same base knowledge for conducting their research. This implies a focus by the group of authors, not only on the topic, but on the paradigm supporting the topic. As such, the papers in the group may be thought of as reporting on research that is concerned with the same Kuhnian puzzle, because the researchers all cite the same exemplar references. Given the importance of research fronts as representative of puzzles and sub-specialties within the specialty covered by the collection of papers, it is very useful to map the relations of the research fronts to other groups of entities within the collection. Many different matrices can be

formed to relate groups of entities in the paper collection using the crossmapping technique discussed in Chapter 12.3.

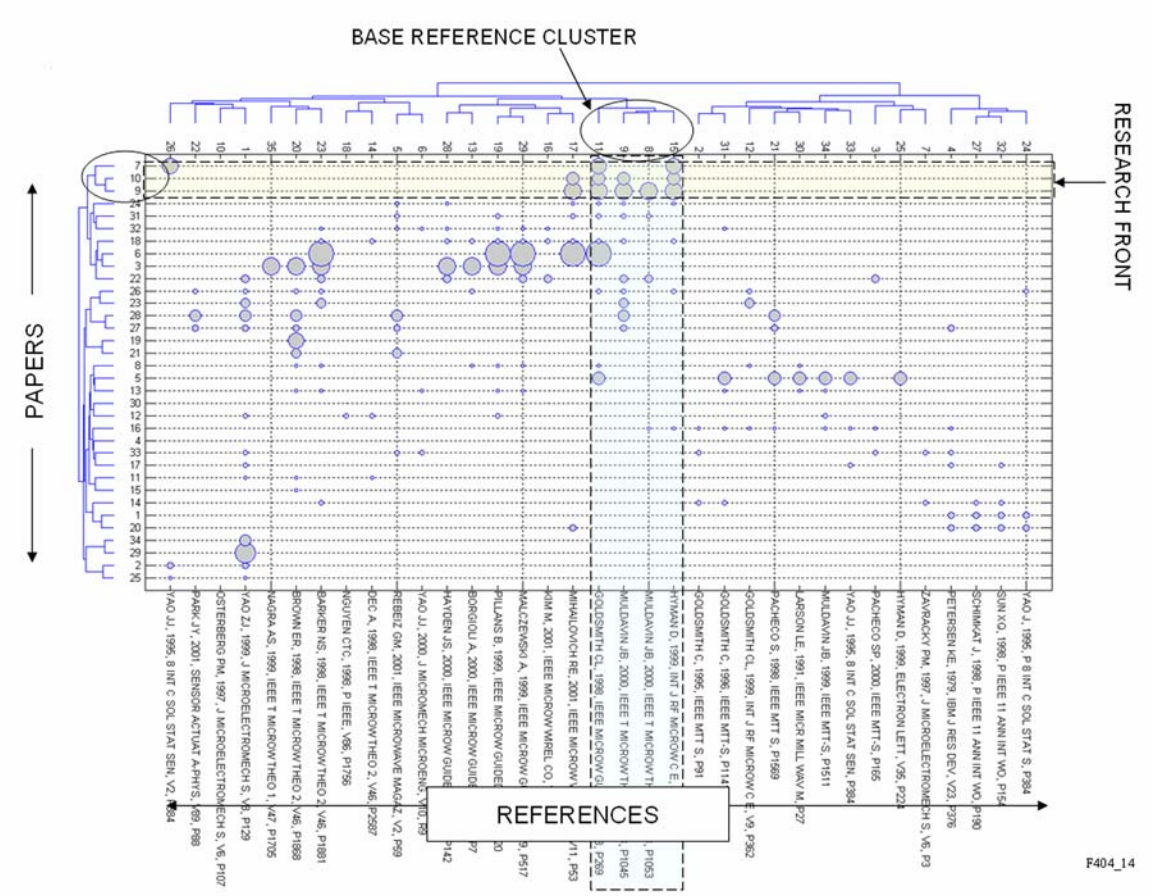


Figure 45. Crossmap of paper groups (clustered using bibliographic coupling) by reference groups (clustered using co-citation) from a collection of papers on the topic of MEMS RF switches. Under Morris and Yen's definition, a reference front is a group of papers, such as the papers in the three groups of highlighted rows, while a base reference group is a co-citation cluster such as the 4 references highlighted in columns. The crossmap shows the correspondence between research fronts and base reference groups as clumps of circles, as can be seen at the intersection of the highlighted rows and columns.

13.5 Author co-citation analysis

Author co-citation analysis was originally introduced by White and Griffith (1981) as a way of mapping the structure of a scientific field. The author co-citation count of a pair of authors is the number of papers that cite both authors. Highly cited reference authors, similar to references, can be thought of as concept symbols of base knowledge, but on a more general level since well-cited authors tend to have many papers on a broad array of topics. A good description of a group of heavily co-cited reference authors then, would be a *school of thought*, as this implies a group of like minded individuals who provide specialized

knowledge to a general audience. Assuming that the heavy citation of reference authors gives evidence of their position as authorities in the specialty, this leads to the interpretation that groups of heavily cited reference authors correspond to groups of experts in the specialty.

The most common method for author co-citation, summarized by McCain (1990) uses a reference author co-citation matrix. As described by McCain, the co-citation matrix is built directly from queries to Dialog, an online database service[2], and does not entail building a collection of papers. This method does not produce feature vectors for the reference authors as described in Chapter 10. Rather, the rows of the co-citation matrix itself are used as feature vectors for clustering and mapping. There is some question about what values to put along the diagonal of the co-citation matrix. McCain discusses the technical problems associated with finding values for the diagonal, but states that usually some scaled value of the number papers citing each author is placed on the diagonal. The co-citation matrix is squared to find values for computation of similarities. White (2003a) advocates the use of correlation coefficient similarity, other researchers (Ahlgren, Jarneving, & Rousseau, 2003) advocate the use of cosine similarity. The resulting similarity measure between two reference authors is not a direct measure of the number of times the reference authors are cited together. The pattern of the number of co-citations with other reference authors is the feature vector of this technique, and so the similarity is based on the similarity of that pattern between pairs of reference authors.

When dealing with a collection of papers, author co-citation analysis is relatively straightforward compared to author co-citation of data acquired through Dialog queries. For a collection of papers, the paper to reference author matrix is built from the paper to reference and reference to reference author matrices.

$$\mathbf{O}[p;ar] = \mathbf{O}[p;r] \cdot \mathbf{O}[r;ar] \tag{112}$$

This results in a non-binary paper to reference author matrix. The co-occurrence matrix can be calculated in two ways. The first method first converts all the non-zero elements of the paper to reference author matrix $\mathbf{O}[p;r]$ to unity before the matrix multiplication:

$$\mathbf{C}[ar;p] = \min(\mathbf{O}[ar;p],1) \cdot \max(\mathbf{O}[p;ar]) \tag{113}$$

This counts the number of papers in which each pair of reference authors appears together. This can also be expressed in terms of a link weight function as:

[2] Thomson-Dialog, 11000 Regency Parkway, Suite 10, Cary, North Carolina, 27516.

$$\mathrm{f}_2\left(o_{ik}[ar;p], o_{kj}[p;ar]\right) = \min(o_{ik}[ar;p],1) \cdot \min(o_{kj}[p;ar],1) \tag{114}$$

and the path combining function f_1 is a summation:

$$\mathrm{f}_1 = \sum_{k=1}^{np} \mathrm{f}_2\left(o_{ik}[ar;p], o_{kj}[p;ar]\right) \tag{115}$$

The second method to count co-occurrences is by the overlap method:

$$\mathbf{C}[ar;p] = OVL\left(\mathbf{O}[ar;p], \mathbf{O}[p;ar]\right) \tag{116}$$

Analysis of the reference author co-citation matrix C[ar;p] is straightforward and similar to analysis of co-citation discussed in Chapter 13.2. Crossmapping techniques can be easily applied to show the relation of reference author groups as schools of thought to the research fronts within the collection of papers.

13.6 Journal co-citation analysis

As discussed by White and McCain (1989), most of the investigations of journals focuses on analyzing information flow among journals based on a "cross citation" matrix. This is a matrix of paper journals to reference journals where the elements (*i*,*j*) count the number of citations from papers in paper journal *i* to references containing reference journal *j*. This type of analysis is called *journal network analysis* in (McCain, 1991) and is facilitated by the wide availability of cross-citation matrices from ISI's Journal Citation Reports. In a collection of papers this cross-citation matrix is easily calculated from the occurrence matrices:

$$\mathbf{O}[jp;jr] = \mathbf{O}[jp,p] \cdot \mathbf{O}[p,r] \cdot \mathbf{O}[r,jr] \tag{117}$$

An example of journal network analysis was conducted by Doreian (1988), where he showed that the journal *positions* of individual journals in the discipline of Geography, derived from clustering based on patterns in the cross-citation matrix, closely matched journal *roles,* which are classifications of journals made by subject matter experts. Starting in the early 1990's, it became possible, using Dialog queries, to extract counts of the number of times that pairs of references journals are cited in individual papers. This allows *journal co-citation analysis*, an analysis method very similar to author co-citation analysis discussed in Chapter 13.5. McCain (1991), introduces this technique. Further examples of journal co-citation analysis can be found in Ding, Gobinda, Chowdury, and Foo (2000), and McCain (1998).

Similar to the discussion on author co-citation analysis, this type of analysis can be carried out using rows of the paper to reference journal matrix as feature vectors or using rows of the journal co-citation matrix as features. The journal co-citation matrix can be calculated from:

$$\mathbf{C}[jr, p] = \min(\mathbf{O}[jr, p],1) \cdot \min(\mathbf{O}[p, jr],1) \tag{118}$$

using the min function to convert from a non-binary paper to reference journal matrix to a binary paper to reference journal matrix. This can be calculated as a link weight function, similar to Equation (115) and (116) for author co-citation analysis. The reference journal co-citation matrix can also be calculated using the overlap function:

$$\mathbf{C}[jr, p] = OVL(\mathbf{O}[jr, p], \mathbf{O}[p, jr]) \tag{119}$$

Groups of reference journals that are clustered using journal co-citation are journals that are supplying base knowledge to common papers. Assuming such papers represent research topics or Kuhnian puzzles, then these groups of reference journals can be considered *base knowledge libraries*, or *base knowledge archives*.

13.7 Braam-Moed-vanRaan (BMV) co-citation co-word analysis

The method of Braam, Moed and van Raan (1991) uses clustering of references and their relations to terms to analyze collections of journal papers. Figure 46 shows a diagram of the BMV analysis method. The method first clusters references into base reference groups based on co-citation. The papers in the collection are assigned to overlapping groups based on the base reference groups they cite. After this, *word profile groups,* consisting of overlapping groups of terms, are formed based on the frequency of terms in the paper groups. This allows relating terms to base reference clusters and assists in labeling base reference groups and searching the paper collection. Assuming that the list of co-citation clusters is in the membership matrix **G**[*gr*,*r*], then the relations from the base reference clusters to terms can be found using:

$$\mathbf{O}[gr;t] = \mathbf{G}[gr,r] \cdot \mathbf{O}[r,p] \cdot \mathbf{O}[p,t] \cdot \mathbf{G}[t,gt] \tag{120}$$

The cascade of relations is noted along the bottom of Figure 46. In this figure, four separate bipartite graphs are noted, with the matrices that represent those graphs shown in the equation below. The base reference group to reference membership matrix **G**[*gr*;*r*] is produced by clustering references by co-citation, while the term to word profile group membership matrix **G**[*t*;*gt*] is compiled indirectly by compiling lists of words in groups of *central* papers, that is, groups of papers that cite references in only one base reference group.

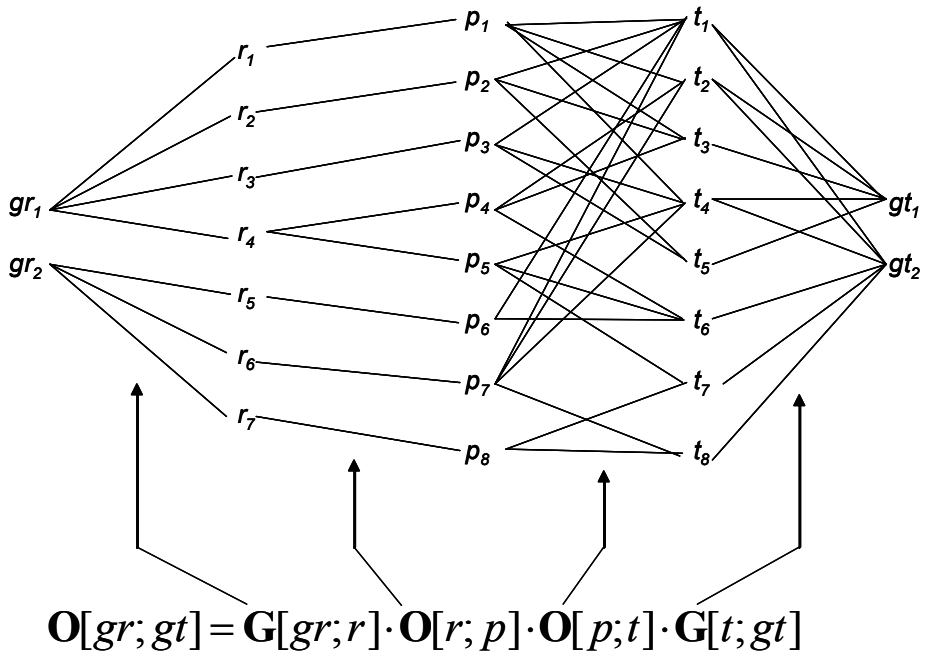


Figure 46. Diagram of the BMV analysis method which relates groups of references to groups of terms (word profile groups).

13.8 Latent semantic analysis

Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), also commonly known as Latent Semantic Indexing, is a latent variable technique that uses singular value decomposition to relate both papers and terms to abstract "concepts." This computationally expensive technique is designed to help disambiguate synonyms, groups of terms that have the same meaning, and polynyms, groups of meanings associated with the same term. Mathematically the technique is easily described as a matrix equation:

$$\mathbf{O}[t; p] = \mathbf{O}[t; c] \cdot \mathbf{O}[c; c] \cdot \mathbf{O}[c; p] \tag{121}$$

The term to paper matrix, $\mathbf{O}[t;p]$, is decomposed, using singular value decomposition (SVD), into the product of $\mathbf{O}[t;c]$, the term to concept matrix, $\mathbf{O}[c;c]$, a diagonal matrix of singular values, (analogous to scale parameters), and $\mathbf{O}[c;p]$, the concept to paper matrix. The singular values in $\mathbf{O}[c;c]$ are ranked by the ability of their corresponding concepts to explain variance in the co-occurrence matrix $\mathbf{C}[t;p]$. In order to avoid overfitting of terms to concepts, and to allow fast execution of queries, the number of concepts, *nc*, is selected to be much less than *nt*, the number of terms, with *nc*, usually fixed at about 300 or so (Deerwester et al., 1990). The reduction is accomplished by applying a threshold to the singular values. The product of

the three matrices on the right side of Equation (102) is an approximation to the term to paper matrix $\mathbf{O}[t;p]$ in the least squares sense. Note that matrices $\mathbf{O}[t;c]$ and $\mathbf{O}[c;p]$ are not strictly occurrence matrices as defined in Chapter 4.1, and both contain negative as well as positive elements.

Given some query vector, $\mathbf{Q}[t]$, that is expressed as a term vector, a query concept vector $\mathbf{X}$ can be calculated (Berry, Dumais, & O'Brien, 1995) as:

$$\mathbf{X}[c] = \mathbf{Q}[t] \cdot \mathbf{O}[t;c] \cdot \mathbf{O}[c;c] \tag{122}$$

This vector can be compared, usually using the cosine measure, to columns of the concept to paper matrix, $\mathbf{O}[c;p]$, to obtain a ranked list of papers associated with the query's concepts. Similarly, example papers can be used to find associated terms and papers that use the same concepts as the example paper.

Updating the SVD, that is, recalculating the three matrices on the left side of Equation (102) when new papers are added is problematic and computationally expensive (Berry et al., 1995). There is no reason that the LSA technique cannot be applied to other occurrence matrices in the collection of papers. For example, SVD decomposition of the reference to paper matrix would allow identification of synonymous references and polynymous meanings of references.

13.9 Hubs and authorities

The hubs and authorities technique is designed for web search engines (Kleinberg, 1999). Assuming a bipartite network consisting of citing and cited entities, authorities are cited entities that receive many citations and are co-cited often with other authorities. Hubs are citing entities that have many outgoing citations that tend to cite the same entities that are cited by other hubs. Highly cited entities not widely cited by hubs are not authorities, while citing entities with many outgoing citations that do not consistently cite authorities are not hubs.

Kleinberg's algorithm for identifying hubs and authorities is an iterative technique that measures the amount of "authority" of all cited entities and measures the "hubness" of all citing entities. Assume for explanatory purposes that the citing entities are papers and that the cited entities are references. The occurrence matrix is $\mathbf{O}[p;r]$ is of dimension np by nr. Now assume an nr by 1 vector $\mathbf{X}[r]$ such the $x_i[r]$ is equal to the magnitude of the "authority" of reference i. The sum of squares of the elements of $\mathbf{X}$ is normalized to be equal to unity. Also assume an np by 1 vector $\mathbf{Y}[p]$ such that $y_i[p]$ is equal to the magnitude of the "hubness" of paper i. The sum of the squares of the elements of $\mathbf{Y}$ is also normalized to be equal to unity. Kleinberg's algorithm is:

Initialize:

$$\mathbf{Y}_1[p] = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \qquad \mathbf{X}_1[r] = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{123}$$

Iterate i:

$$\mathbf{Y}_{i+1}[p] = \mathbf{O}[p;r] \cdot \mathbf{X}_i[r] \tag{124}$$

$$\mathbf{X}_{i+1}[r] = \mathbf{O}[r;p] \cdot \mathbf{Y}_i[p] \tag{125}$$

normalize $\mathbf{X}_{i+1}[r]$ *and* $\mathbf{Y}_{i+1}[p]$

$$norm(\mathbf{X}_{i+1}[r]) : (x_{i+1}[r])_j = \frac{(x_{i+1}[r])_j}{\sqrt{\sum_{k=1}^{nr} (x_{i+1})_k}} \tag{126}$$

$$norm(\mathbf{Y}_{i+1}[p]) : (y_{i+1}[p])_j = \frac{(y_{i+1}[p])_j}{\sqrt{\sum_{k=1}^{np} (y_{i+1}[p])_k}} \tag{127}$$

Next i

Kleinberg has shown that after many iterations $\mathbf{X}[r]$ will asymptotically approach the vector of the eigenvalues of $\mathbf{C}[r;p]$, the co-citation matrix, while $\mathbf{Y}[p]$ will approach the vector of the eigenvalues of $\mathbf{C}[p;r]$, the bibliographic coupling matrix. For collections of web pages, Kleinberg reports that convergence occurs quickly, with as few as 20 iterations. A threshold can be applied to the elements of **X** and **Y** to distinguish authorities among the references and hubs among the papers respectively.

Although this algorithm has been applied to collections of web pages, it is possible to conjecture on the interpretation of hubs and authorities when the technique is applied to collections of papers. Authorities can be interpreted as exemplar references in the specialty, since they correspond to references that are

consistently cited in groups. Hubs should correspond to review papers, since the cite a large number of exemplar references in a specialty. Another interpretation of hub papers could be that they correspond to "puzzle-solving" papers, i.e., papers working within the Kuhnian paradigm of the specialty, and citing a "standard" set of "authorities", i.e., exemplar references in a sub-specialty.

13.10 Author identities and author images

The concept of author identity and author image was introduced by White (White, 2001). Given a paper author, that author's identity is the list of reference authors that are cited by that author. In the mathematical treatment presented here, for paper author *i*, the author's identity is the feature vector $\mathbf{O}_i[ap,ar]$. that is, row *i* of the paper author to reference author matrix. Given a reference author *j* the author's image is the list of reference authors that reference author *j* has been co-cited with. This can be represented as $\mathbf{C}_i[ar,ap]$, which is row *i* (or column *i*) on the reference author co-citation by paper author matrix. The author identity characterizes a paper author by the pattern of reference authors the paper author cites. This identifies the author by the school of thought upon which his/her work is based. The author image characterizes a reference author by the pattern of reference authors with which he/she tends to be co-cited. This identifies the school of thought to which the reference author belongs.

The concept of author images and identities can be generalized to the other two types of feature vectors in a paper author to reference author matrix. Figure 47 shows an example paper author to reference author matrix and its associated co-occurrence matrices. Given a paper author *i*, as noted in the figure, the vector highlighted in blue represents the author identity, which has been redesignated as *reference author identity of a paper author*. A second paper author feature vector, row *i* in the paper author coupling matrix, **C**[*ap*;*ar*], is designated as *paper author identity of a paper author*. This is the list of paper authors that tend to cite the same reference authors as paper author *i*. Given a reference author *j*, the author image, designated *reference author image of a reference author*, is column *j* (or row *j*) of the reference author co-citation matrix **C**[ar;ap]. A second reference author feature vector, designated *paper author image of a reference author*, is column *j* of the paper author to reference author matrix **O**[*ap*,*ar*], which is, of course a row of the reference author to paper author matrix **O**[*ar*,*ap*]. This vector is a list of the paper authors that have cited reference author *j*. Given these 4 feature vectors, there are four ways to characterize a physical author in terms of other physical authors:

- **paper author image**: the authors who cite the author of interest. These are the authors that draw upon the author of interest for base knowledge, an author's *knowledge sinks*.

- **reference author image**: the authors who are cited together with the author of interest. These are the authors supplying base knowledge to the same authors as the author of interest, an author's *knowledge co-sources*.

- **reference author identity**: the authors that the author of interest cites. These are the authors that the author of interest draws upon for knowledge, an author's *knowledge sources*.

- **paper author identity**: the authors that cite the same authors as the author of interest. These are the authors that use the same knowledge sources as the author of interest, an author's *knowledge co-sinks*.

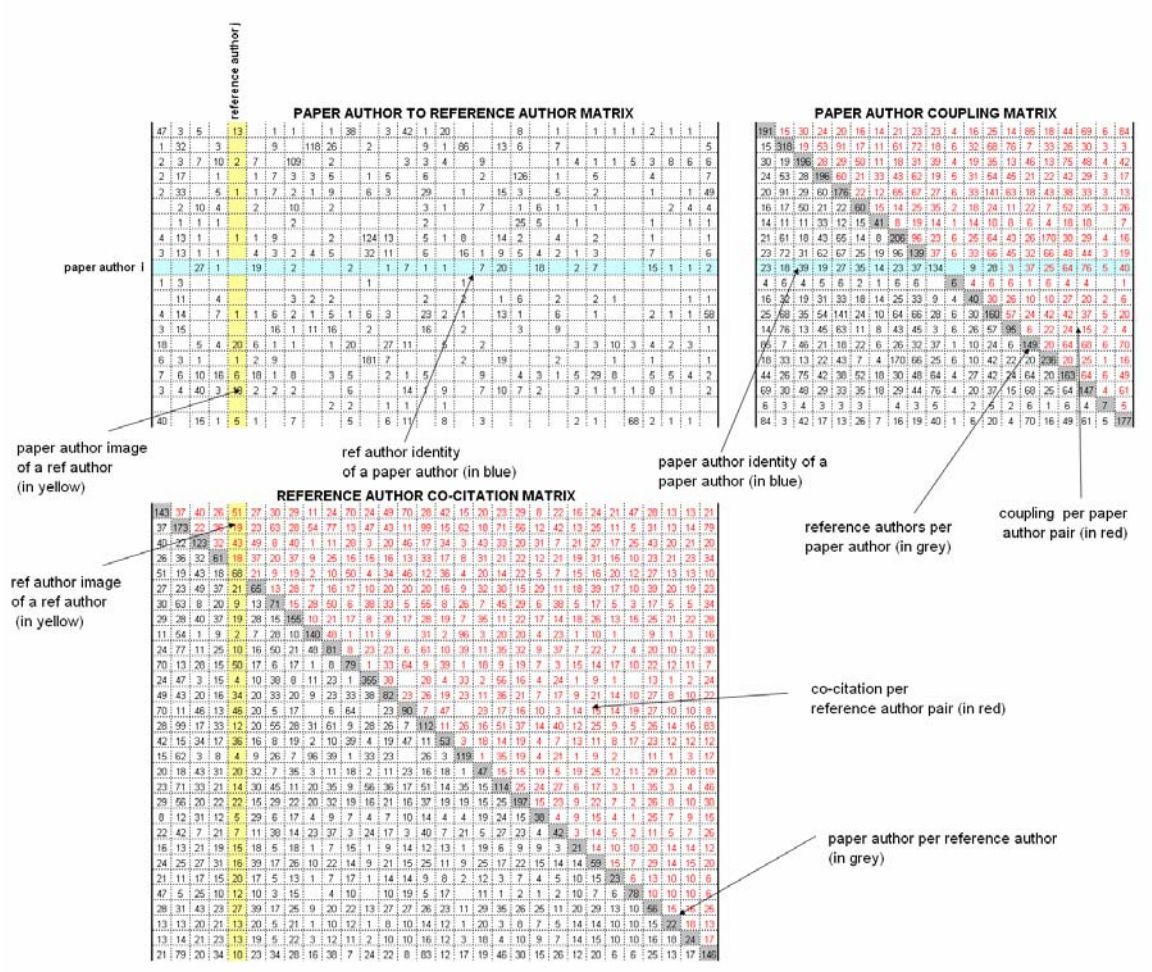


Figure 47. Generalization of author images and identities for reference authors and paper authors related by a paper author to reference author matrix.

13.11 Pathfinder networks

Pathfinder network analysis is a link pruning technique used to visualize structure in networks (Schvaneveldt, Durso, & Dearholt, 1989). This type of analysis was applied by Chen (1998) to visualizing collections of papers. As originally proposed (Schvaneveldt, Dearholt, & Durso, 1988), the method expresses link weights in networks as *distances* whose magnitude is proportional to the *weakness* of the connection between those nodes. In this sense, two identical nodes are connected by a link of weight zero, while for two totally unrelated nodes the link weight is infinity. This is opposite of the convention adopted in this report, that is, the definition given in Chapter 3.1, where link weights are proportional to the strength of the connection between nodes those links connect. To maintain uniformity, the discussion of pathfinder analysis will be adapted here to links defined as having weight proportional to strength of connection, in accordance with the definition used throughout this report. Assume a weighted, undirected graph whose link weights are given by a co-occurrence matrix $\mathbf{C}[x_1;x_2]$. A similarity matrix can also be used for this analysis. A *pfnet*, **P**, of this network is a pruning where the "weak" links in the network are dropped. This allows visualization of the dominant structure of the network and the principle channels of communications among the nodes. The pruning algorithm has parameters r and q, and can be summarized as follows:

- Save the original co-occurrence matrix, $\mathbf{C}[x_1;x_2]$, as $\mathbf{C}^1$.
- A second matrix $\mathbf{C}^q$ is computed by using the inverse Minkowski metric, with parameter r, to measure the weight c^q_{ij} of each of the possible paths from node i to node j that are of path length q or less. The matrix element c^q_{ij} of $\mathbf{C}^q$ is the weight of the maximum weight path from node i to node j, and can be considered the "path of least resistance" between those nodes. The inverse Minkowski metric is used for path weight calculations
- The pfnet, a pruned matrix **P,** is computed by comparing $\mathbf{C}^q$ to $\mathbf{C}^1$. If c^q_{ij} is less than c^1_{ij} then $p_{ij} = 0$, otherwise $p_{ij} = c^1_{ij}$.

The effect of the pruning is that all links in the network that are "short-circuited" by a path of greater weight are dropped from the network. This pruning produces a "backbone" structure that, when visualized, helps to understand the structure of the network. Because of the need to find all the possible paths in the network of path length q or less, in practice the actual computation of the pathfinder network is somewhat complicated and computationally intensive. In matrix terms a straightforward way to compute the matrix $\mathbf{C}^q$ is to iteratively apply cascaded bipartite network link weight calculations.

As discussed in Chapter 3.2 define a link weight function where:

$$\mathrm{f}_2(i,j,k) = \left[\left(c^m_{ik}\right)^{-r} + \left(c^m_{kj}\right)^{-r}\right]^{\frac{1}{r}} \tag{128}$$

and

$$\mathrm{f}_1(i,j) = \max\left[\mathrm{f}_2(i,j,1), \mathrm{f}_2(i,j,2), \ldots, \mathrm{f}_2(i,j,nx_1)\right] \tag{129}$$

This defines a path weight function PATHW:

$$\mathbf{C}^{m+1} = \mathrm{PATHW}(\mathbf{C}^m, \mathbf{C}^1, r):$$

$$\mathrm{c}_{\mathrm{ij}}^{\mathrm{m+1}} = \max\left(\left[\left(c_{i1}^m\right)^{-r} + \left(c_{1j}^1\right)^{-r} \right]^{-\frac{1}{r}}, \ldots \left[\left(c_{inx_1}^m\right)^{-r} + \left(c_{nx_1 j}^1\right)^{-r} \right]^{-\frac{1}{r}} \right) \tag{130}$$

Using the path weight function, the pathfinder network algorithm is solved iteratively:

```
Initialize P = C^1
for m = 1 to q-1:
```

$$\mathbf{C}^{m+1} = \mathrm{PATHW}(\mathbf{C}^m, \mathbf{C}^1): \tag{131}$$

```
    %set P elements to zero if corresponding C^(m+1) elements are greater
```

$$p_{ij} = \begin{cases} 0 & if\ p_{ij} < c_{ij}^m \\ p_{ij} & otherwise \end{cases}, \text{ for all i,j} \tag{132}$$

```
repeat
```

After iteration, the matrix **P** is the adjacency matrix of the pfnet. The algorithm is easily programmed and is readily adaptable to sparse matrix techniques for rapid calculation. Note that for the case of *r* equal infinity, then the path weight function of Equation (130) reverts to:

$$\mathbf{C}^{m+1} = \mathrm{PATHW}(\mathbf{C}^m, \mathbf{C}^1, \infty):$$

$$\mathrm{c}_{\mathrm{ij}}^{\mathrm{m+1}} = \max\left[\min\left(c_{i1}^m, c_{1j}^1\right), \ldots \min\left(c_{i\ nx_1}^m, c_{nx_1\ j}^1\right) \right] \tag{133}$$

Pathfinder analysis can be applied to any similarity matrix in a collection of journal papers. For example, Chen, Cribbin, et al, (2002) used pathfinder analysis to study graphs of references based on similarity derived from co-citation. White (2003b), similarly used pathfinder analysis to examine networks of reference authors using similarity based on author co-citation counts.

14. SOFTWARE TOOLKIT

14.1 Introduction

This chapter discusses the software toolkit developed for analysis and visualization of data from collections of papers and patents. The creation of this software was motivated by the need for a visualization tool for exploring and mapping collections of papers according to research subtopics. Importantly, the software has provided a testbed for experimentation in data storage, analysis, and visualization techniques when dealing with collections of papers.

The toolkit was first conceived as a means for analysis and visualization of collections of patents and of papers gathered from abstract services. An early version of the software was reported in the literature by Morris, DeYong, Wu, Salman and Yemenu. (2002). In that version of the toolkit, called DIVA, for Database Information Visualization and Analysis, visualization was done using two-dimensional multi-dimensional scaling (MDS) maps, which were produced using VxOrd, a utility of Sandia's VxInsight software (Boyack, Wylie, & Davidson, 2002). Alternately, mapping was performed using a method based on self-organizing map (SOM) neural networks (Morris, Wu, & Yen, 2001). In this early version of the software, the data structures were built around the visualizations, with no realization of the cascaded bipartite structures that have been discussed in this report.

In its current version, the toolkit incorporates the matrix-based data structures that are explained in Chapters 4 and 5. Additionally, the software incorporates the visualization techniques described in Chapter 12 and allows many of those visualizations to be realized as web pages for dissemination to groups of subject matter experts. The software can produce interactive timeline webpages that allow remote subject matter experts to execute database queries to produce useful summary data of clusters in the collection.

The software is mostly operated through a graphical user interface (GUI). It is built around six sets of routines:

- **Main GUI routines**: these main routines maintain the data structure and provide a user interface.
- **Data input routines**: these routines load tables from the paper collection database into matrices in MATLAB variable space.

- **Clustering routines**: these routines cluster entities based on co-occurrence.
- **Mapping routines**: these routines produce timelines, crossmaps and usage maps and allow users to interact with those visualizations.
- **Plotting routines**: these routines allow plotting of several types of distributions taken from the collection.
- **Report generating routines**: these routines produce reports on groups of entities derived through clustering.

Discussions of the toolkit will start with a description of the main GUI and its features, a description given in the context of conducting a case study for some practical purpose such as technology forecasting.

14.2 General use of toolkit software

Figure 48 shows a diagram of a typical sequence of tasks that a user follows to perform a study of a specialty through its journal literature. The user typically executes a series of queries in an iterative fashion to build a collection of papers that covers the specialty of interest. These queries will produce a series of text files that contain a sequential collection of records corresponding to the data from individual papers. These text records are pulled into an MS ACCESS database using a Visual Basic program module in the database. Once the records are in the database, the data is loaded into MATLAB matrix variables through an ODBC link. At this point the toolkit software can be used to cluster various entities such as papers, references, and paper authors using similarities calculated through co-occurrence counts, as discussed in Chapter 11. At this point the user will produce several visualizations and interactively explore them in order to seek information about the specialty. The user may print out those visualizations for inclusion in a report, or may post the visualizations on a project website for use of subject matter experts. The user can also plot indicators of the state of the specialty, usually plots of distributions as discussed in Chapter 6. The user may also produce text reports that can be presented in tabular form to subject matter experts.

Usually, the user is looking to identify key entities in the collection, e.g., highly productive authors, or highly cited references. The user also desires to cluster entities into meaningful groups. For example, papers should be clustered in groups by topic, references can be clustered into groups that tend to be cited together, paper authors can be clustered to show teams of collaborators. It is also desired to show the relation of groups to one another, for example, using a dendrogram, or additionally, the user may want to show the overlapping relationships of groups of two different entity-types.

The remainder of this chapter will describe the collections of routines in the toolbox software that are used to facilitate the performance of the tasks shown in Figure 48.

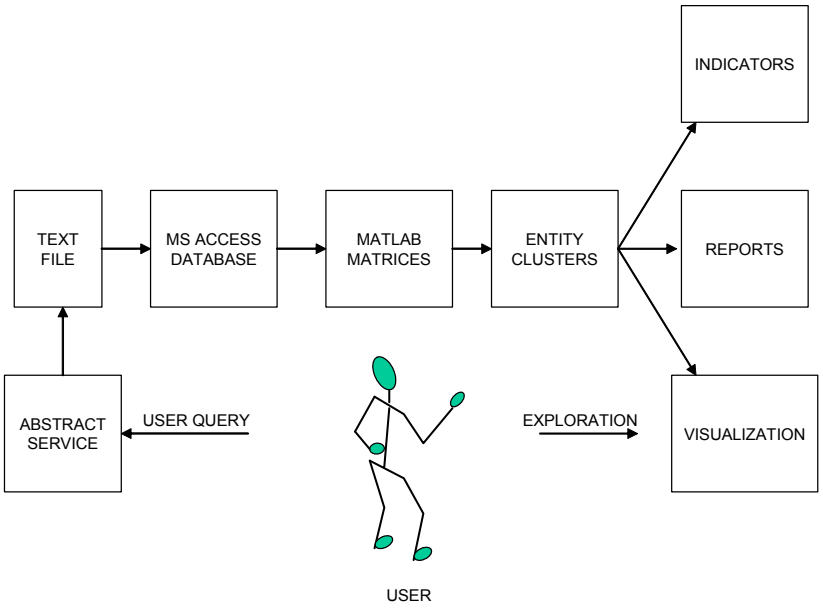


Figure 48. Diagram of typical sequence of steps when conducting a case study to investigate a specialty through exploration of its literature.

14.3 Database tables

Web of Science source files. Data is acquired in the form of collections of papers or patents. For brevity, this report will concentrate on collections of papers gathered from ISI's Web of Science product, but this discussion can be easily generalized to cover patents. Assuming a Web of Science collection, the data is assumed to be topic-specific concerning some scientific specialty, and is gathered either by executing queries on search terms or by finding sets of papers that cite seed references. Seed references are important references in the specialty that are assumed to be cited by most papers in the specialty. In actual practice, the acquisition of a collection of papers that covers a specialty well is an iterative process that starts with executing a well-constructed set of queries, finding the appropriate seed references from the resulting gathered set of papers, and then collecting additional papers that cite those seed references.

The data, when downloaded, is in the form of records that correspond to journal papers. Each record contains the following information:

- Paper title
- Paper authors
- Journal name
- Journal volume
- Journal issue
- Journal page number

- Journal year
- Abstract
- Author index terms
- ISI generated index terms
- Cited references

Each of the cited references contains the following information:

- Reference first author name
- Reference year
- Reference journal
- Reference volume
- Reference first page

If references correspond to books, the reference information consists of the first author, book title and year. Other types of references, such as films, web pages, and electronic archives, appear to be handled on an ad hoc basis. Table 5 shows an example record from a Web of Science file.

Database loading routine: An MS VBA (Visual Basic Applications) program, running as a module in a template MS ACCESS database is used to load the data from Web of Science tagged files into MS ACCESS database tables. The program, READ_DATA, is approximately 500 lines in length. It creates and populates 6 tables: 1) the working table of papers and their attributes, 2) citation table, 3) paper author table, 4) abstract table, 5) institution table, and 6) reference key table. The routine can read multiple input files, discards duplicate records, and produces a log file for auditing purposes.

Basic database structure. The basic structure of a database holding a report-based information structure is a series of relational tables. Some of these tables hold the index keys and simple attributes of the entities. Other tables list the associations between entities themselves.

Index tables. Index tables contain a list of all entities of a particular entity-type and their index keys. For example an index table of journal papers will contain an entry for each paper in the collection. Index tables additionally will contain attributes that are associated with each entity in the table. For example an index table of journal papers may contain the title and date of publication of each paper in the table as attributes.

Index tables may also contain associations to entities of other entity-types if only one entity of that entity-type can be associated with the table's index entity. For example, journal papers can only be associated

with one publishing journal, and so the association of each paper with its publishing journal can be listed in the index table for papers. The following index tables are often used by the toolkit:

- ***Working_table***. The working table is the basic table of records in the database. Each record in this table contains a paper ID number which is used to associate all other data in the database. The table also contains the title, paper journal, volume, issue, page number, and date of publication.
- ***Cite_keys table***. This table contains a list of references and their keys, their associated reference journal and reference author, and reference year.
- **Other index tables.** *Author_keys*, *ref_author_keys*, and *id_keys*, are simple index tables for paper authors, reference authors, and index terms respectively.

Table 5. Example record in ISI tagged file format.

INFORMATION TYPE	EXAMPLE ENTRY
PUBLICATION TYPE:	PT J
AUTHOR:	AU Liu, WL
	Lu, Y
	Chen, YH
TITLE:	TI Bioinformatics analysis of SARS-Cov M protein provides information for vaccine development
SOURCE (JOURNAL):	SO PROGRESS IN NATURAL SCIENCE
LANGUAGE	LA English
DOCUMENT TYPE:	DT Article
AUTHOR KEYWORDS:	DE SARS-Cov, M protein, anti-SARS vaccine strategy
KEYWORDS PLUS:	ID TRANSMISSIBLE GASTROENTERITIS CORONAVIRUS; ACUTE RESPIRATORY SYNDROME; VIRUS; IDENTIFICATION; ANTIBODIES; REGION
ABSTRACT:	AB The pathogen causing severe acute respiratory syndrome (SARS) is identified to be SARS-Cov. It is urgent to know more about SARS-Cov for developing an efficient SARS vaccine to prevent this epidemic disease. In this report, the homology of SARS-Cov M protein to other members of coronavirus is illustrated, and all amino acid changes in both S and M proteins among all available SARS-Cov isolates in GenBank are described. Furthermore, one topological trans-membrane secondary structure model of M protein is proposed, which is corresponded well with the accepted topology model of M proteins of other members of coronavirus. Hydrophilic profile analysis indicated that one region (aa150similar to210) on the cytoplasmic domain is fairly hydrophilic, suggesting its property of antigenicity. Based on the fact that cytoplasmic domain of the M protein of some other coronavirus could induce protective activities against virus infection, this region might be one potential target for SARS vaccine development.
RESEARCH ADDRESSES:	C1 Tsing Hua Univ, Immunol Lab, Biomed Sci Res Ctr, Beijing 100084, Peoples R China.
	Tsing Hua Univ, Dept Biochem, Beijing 100084, Peoples R China.
	MOE, Prot Sci Lab, Beijing 100084, Peoples R China.
	RP Chen, YH, Tsing Hua Univ, Immunol Lab, Biomed Sci Res Ctr, Beijing 100084, Peoples R China
CITED REFERENCES:	CR ANTON IM, 1996, VIRUS RES, V46, P111
	DROSTEN C, 2003, NEW ENGL J MED, V348, P1967
	GABRIELLA E, 2003, J VIROL METHODS, V109, P139
	KROGH A, 2001, J MOL BIOL, V305, P567
	KYTE J, 1982, J MOL BIOL, V157, P105
	LIU WL, 2003, FEMS IMMUNOL MED MIC, V35, P141
	LU Y, 2003, CHINESE SCI BULL, V48, P1115
	MARCO AM, 2003, SCIENCE, V300, P1399
	PEIRIS JSM, 2003, LANCET, V361, P1319
	RISCO C, 1995, J VIROL, V69, P5269
	ROST B, 1994, J MOL BIOL, V235, P13
	RUAN YJ, 2003, LANCET, V361, P1779
	STEPHEN FA, 1997, NUCLEIC ACIDS RES, V25, P3389
	TOOZE SA, 1986, J VIROL, V60, P928
	XIAO Y, 2001, IMMUNOL LETT, V77, P3
CITED REFERENCE COUNT:	NR 15
TIMES CITED:	TC 0
PUBLISHER:	PU TAYLOR & FRANCIS LTD
PUBLISHER CITY:	PI ABINGDON
PUBLISHER ADDRESS:	PA 4 PARK SQUARE, MILTON PARK, ABINGDON OX14 4RN, OXON, ENGLAND
ISSN:	SN 1002-0071
JOURNAL ABBREVIATION:	J9 PROG NAT SCI
O JOURNAL ABBREVIATION:	JI Prog. Nat. Sci.
PUBLICATION YEAR:	PY 2003
PUBLICATION DATE:	PD NOV
VOLUME:	VL 13
ISSUE:	IS 11
BEGINNING PAGE:	BP 844
ENDING PAGE:	EP 847
PAGE COUNT:	PG 4
ISI NUMBER:	GA 764PP
ISI DOCUMENT IDENTIFIER:	UT ISI:000188215500008
END OF RECORD:	ER

Link tables. Link tables are used to list the associations between entities drawn from a pair of entity-types in the collection. In other words, link tables are a list of the links in one of the collection's bipartite networks. Link tables are in "normal form," that is, there is only one entry for each link, even though an entity may have links to several entities of that particular entity-type. Links are often links of association only. This requires only that a table entry for each link contain the index keys of the entities that are associated. For example, a table entry of "55, 39" may denote that paper 55 is associated with author 39. Such links are *unweighted*. Other types of links may have weights associated with them. For example, links between papers and linguistic terms may be weighted by the count of the number of times that the terms occur in the paper. So, for example, an entry of "20, 45, 6" could denote that paper 20 contains 6 occurrences of term 45. The following tables are often created and used by the toolkit:

- **Author table**. This table contains a list of paper authors and their associated papers. Each record associates an author with a paper. For each paper there is a record for each author of the paper.
- **Citation table**. This table contains a list of papers and the references they cite. Each record associates a reference with a paper. For each paper there is one record for each reference cited by the paper.
- **Index terms**. This table contains a list of papers and their associated index terms. Each record associates an index term with a paper. For each paper there is one record for each index term associated with the paper.
- **Abstracts**: This table contains the abstracts of each paper. These are stored line by line.
- **Institution**: This table holds the author institutions associated with each paper. Because of the difficulties of disambiguating different version of institution addresses and also because it is not possible to match institutions to specific authors in multiple author papers, this table is seldom used.

14.4 DIVA main GUI

Assume that the purpose of a study is to investigate a specialty, and that a collection of papers is gathered from ISI's Web of Science product and loaded into an MS ACCESS database. It is at this point that the user starts using the toolkit and will work from the main GUI (graphical user interface) shown in Figure 49.

The main GUI contains four sections:

- **Map section**: this section allows display of maps in several formats. Dotmaps are either two dimensional MDS maps or timelines. These dotmaps can be presented in a density format as surface landscape maps or as contour maps.
- **Connections section**: this section allows display of links on dotmaps. Links are stored as similarity matrices or as directed adjacency matrices. It is possible to display directed and

undirected links to and from a group of papers, and additionally, the dependent or precedent links through citations can be displayed.

- **Clusters section**: this section allows manipulating groups of papers that are highlighted on the map. Groups are identified by queries or by highlighting them by dragging a mouse pointer on a map.
- **Time section**: this section allows altering the display of papers on the map to show their publication date. This can be done by varying the color of dots on the map as a function of publication date, or it can also be used to highlight specific time intervals. These features are useful for showing time relations of papers on MDS maps.

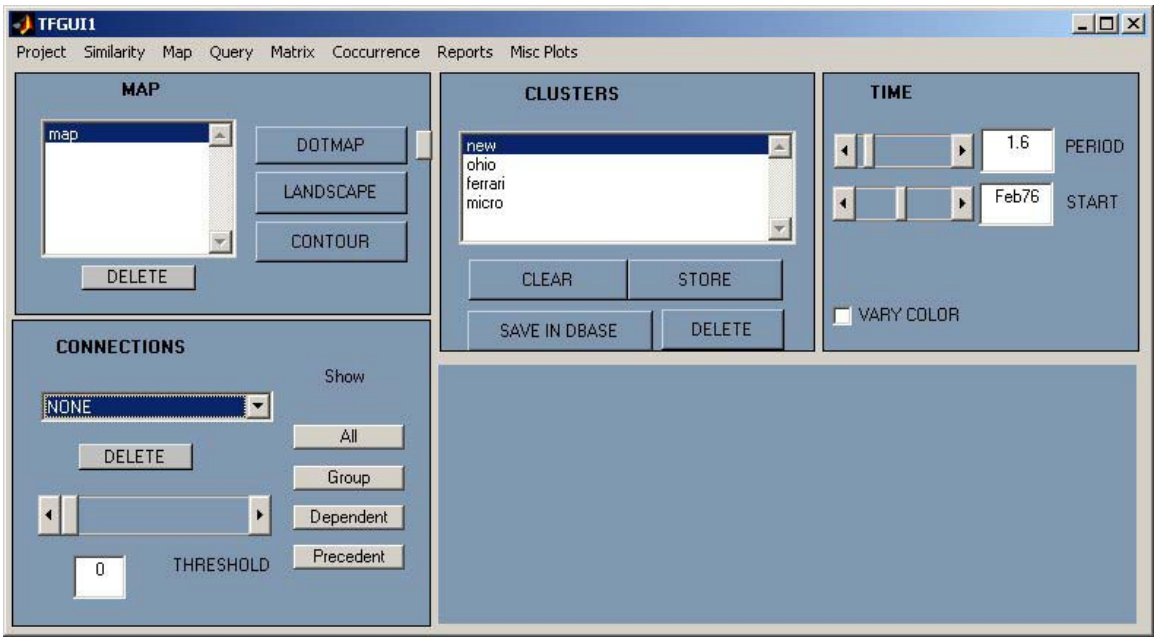


Figure 49. DIVA main GUI.

The main GUI also contains 8 drop down menus along the top of the GUI as shown in Table 6. The *project menu* is used for general project management: setting up new projects before loading data, saving projects, loading projects and setting up the links to the project database. The *similarity menu* is used for loading the adjacency matrix of papers linked by citation. The *map menu* is used to make two dimensional MDS maps. The *query menu* is used to execute queries and highlight them on maps that are being displayed. The *matrix menu* is used to load data from the database in the form of occurrence matrices. The *co-occurrence menu* is used for clustering entities into groups by co-occurrence and making crossmaps of those maps of groups from pairs of entity-types. The *reports menu* is used to produce written reports in rich text format of the characteristics of groups of different entity-types. Finally, there is the *misc plots* menu for making plots of distributions within the collection.

Table 6. Menu items for the main GUI of the toolkit.

MAIN MENU SUBMENUS			
Project: • New project • Open project • Save project • Save project as • Set database link • Convert patent set	**Similarity**: • Read adjacency matrix from database	**Map**: • Make SOM 2D ordination • PFNET GUI	**Query**: • Highlight dots from query
Matrix: • Load paper to reference matrix • Bib to DIVA • Load paper to paper author matrix • Load paper to reference author matrix • Load DE terms matrix • Load assignee matrix	**Co-occurrence**: • Use bibliographic coupling default • Use co-citation default • Use paper author co-occurrence default • Use author co-citation default • Make crossmaps	**Reports**: • Research fronts • Co-citation clusters • Authors/inventors • Institutions	**Misc Plots**: • Cites per paper distribution • Co-citation distribution • Bibliographic coupling distribution • Bibliographic coupling clustering distribution • Reference frequency distribution

14.5 Data input routines

As discussed in Chapter 5, the data in this software is stored as matrices in memory. These are stored as sparse matrices to take advantage of MATLAB's many sparse matrix routines. Table 7 shows a list of the input routines of DIVA that are used to bring data from the database into the memory.

In a typical case study, after gathering the data from the Web of Science and loading it into a database, the first analysis is done by constructing a paper to reference matrix. This is used to cluster papers using bibliographic coupling and to cluster references using co-citation. These two types of analyses allow some exploration of the data collection to assess how well the collection covers the specialty and also whether the collection has enough coherence to be able to map its structure. In many cases, especially in non-technical fields, there is not enough interconnection of entities in the collection to cluster those entities into definite groups. After loading the paper to reference matrix and performing initial exploration of the paper collection, it is useful to load the paper to paper author matrix and the paper to reference author matrix. This allows analysis of author teams by clustering paper authors by co-authorship and assessing experts using author co-citation analysis.

Table 7. List of data input routines.

ROUTINE	DESCRIPTION
ASSIGNEE_MATRIX:	Routine to build patent-assignee matrix
GET_A_P_R	Gets the paper-reference correspondence list
ID_MATRIX	Loads paper to paper id term matrix into DIVA
MAKE_R_YP_MATRIX	Make reference by citing paper year matrix
MAKE_R_YR_MATRIX	Make reference by reference year matrix
P_AP_MATRIX	Loads paper to paper author matrix into DIVA
P_AR_MATRIX	Get paper to ref author matrix
p_de_MATRIX	Loads paper to paper de term matrix into DIVA
P_JR_MATRIX	Get paper to ref journal matrix
P_PJ_MATRIX	Loads paper to paper journal matrix
P_TERM1_MATRIX	Routine to get 1 word terms from database
R_YR_MATRIX	Make reference by reference year matrix
REF_MATRIX	Gets paper-ref matrix from dbase
REF_MATRIX_BY_PAPER	Put an occurrence matrix in lower triangular form

14.6 Clustering routines

Table 8 shows a list of clustering routines in the software. The software is built around a single clustering routine, COOCCUR, which takes a matrix, calculates similarities using a specified similarity metric, performs clustering using hierarchical clustering with Ward's method linkage, and stores the results in a standardized data structure for clustering results. The selection of primary and relative entities, and other parameters for clustering are selected using a GUI, shown in Figure 50.

Table 8. Clustering routines.

ROUTINE	DESCRIPTION
COOCCUR	Routine to cluster row items from matrix
COOCUR_GUI	M-file for coocur_gui.fig
LINKAGE_SIM	MATLAB clustering routine modified to use similarity
PUT_CLUSTERS_AUT	Put author clusters in the database
PUT_CLUSTERS_BIB	Put bib coupling clusters in the database
PUT_CLUSTERS_COC	Put co-citation clusters in the database
PUT_CLUSTERS_WORD1	Puts 1 word term clusters in the database
SEMINAL	Get a list of indexes of key papers

On this GUI two sets of radio buttons are used to select the primary and relative entity-types. As shown, there are 16 possible combinations of primary and relative entity-types, of which 12 are valid. However, only four pairs of entity-types are recognized by the GUI, all other combinations are ignored. These pairs are shown in Table 9.

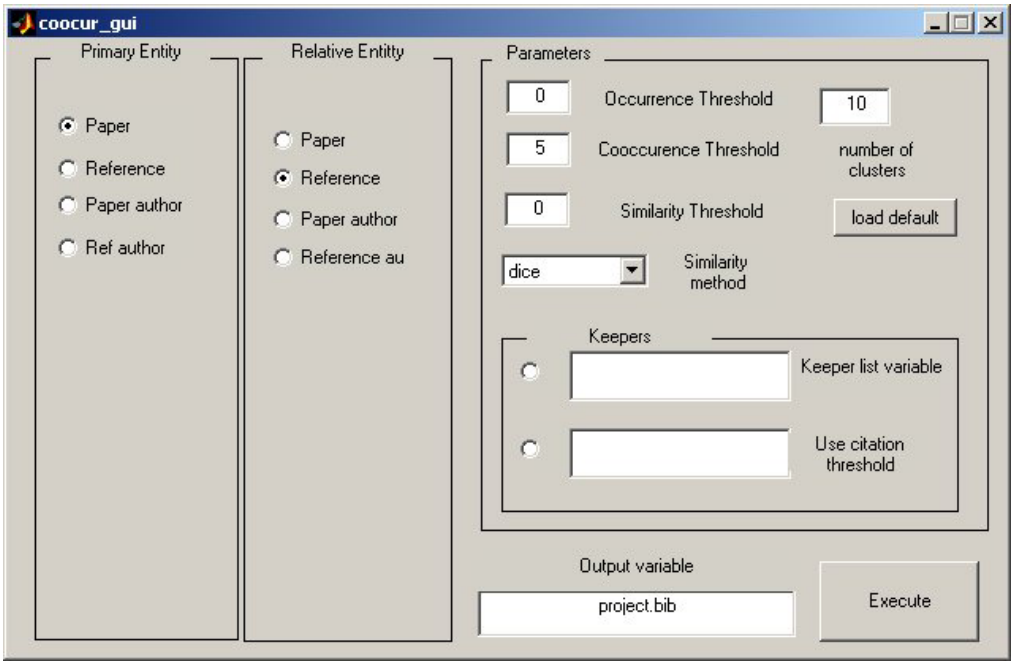


Figure 50. GUI used for setting up and performing clustering.

The clustering parameters are entered on the right side of the GUI. There are parameters for the number of clusters, the type of similarity metric, and the output variable for the cluster structure. The occurrence threshold is used for entities that can occur more than once in the collection, such as references, paper authors and reference authors. This threshold is used to eliminate primary entities that are poorly linked. This can be used to limit clustering to highly cited references or reference authors, or highly productive paper authors, and greatly reduces noise in the clustering. The co-occurrence threshold is used to remove primary entities that do not have many links to other like entities in the collection. The use of this threshold greatly reduces noise in clustering and helps produce coherent clusters with entities that are well related to one another. The similarity threshold eliminates primary entities that do not have some minimum similarity link to at least one other entity. Similar to the occurrence threshold, the similarity threshold reduces noise in clustering by eliminating poorly linked entities.

Table 9. Primary to relative entity-type pairs recognized by the clustering GUI.

Primary entity-type	Relative entity-type	Purpose
paper	reference	Cluster papers using bibliographic coupling
reference	paper	Cluster references using co-citation
paper authors	papers	Cluster paper authors by coauthorship
reference authors	papers	Cluster reference authors by author co-citation

Two "keeper" parameters are available on this GUI. These are provided as a way of including papers in the collections that correspond to highly cited references. These papers often occur early in the collection, and as such, usually have poor bibliographic coupling linkage to other papers in the collection. The citation

threshold parameters force the routine to keep papers that don't meet the occurrence and co-occurrence thresholds but whose corresponding references are well-cited. Alternately, the user can supply a variable name to a vector carrying the indexes of keeper papers that must be retained even if they don't meet the occurrence and co-occurrence thresholds.

After clustering, the results are placed into a structure for later use by mapping routines, particularly, routines for dendrogram seriation. This structure contains seven fields:

- **Call**: this is a substructure which lists the values of the parameters used for clustering. This field is useful for auditing purposes.
- **Z**: this field contains the full cluster tree in the format produced by MATLAB's hierarchical agglomerative clustering routine.
- **Cluster**: this field contains the cluster number of each of the clustered entities.
- **Member**s: this field contains a list of the entities that were clustered. Entities that were discarded using thresholds are not included in this list.
- **Sim**: this field contains the similarity matrix for the clustered entities. This matrix is used by the dendrogram seriation routine.
- **Z1**: this field contains the truncated cluster tree used to find the user requested number of clusters. This tree is used to produce dendrograms when required.
- **Order**: this field contains the serial order of the clusters on the dendrogram. This order is manipulated by the dendrogram seriation routine.

Note that there are also routines in this group for performing similarity calculations using the overlap function or the inverse Minkowski function as discussed in Chapter 3.

14.7 Mapping routines

Table 10 gives a list of the mapping routines developed in the software. These routines can be divided into those routines for interactive exploration, and those used to create web pages. Maps can be divided into timelines and crossmaps.

For timelines, a useful routine is FREQMOD, which makes the size of the dots on the map proportional to the number of times a paper's corresponding reference has been cited. The routine MAKEDENDRO constructs a dendrogram for denoting cluster relations on the map. Another routine, TSPDENDRO, performs dendrogram seriation according to the method of Morris, Asnake, and Yen (2003). The routine TMAP is the main routine for drawing timelines, while the routine MAP_CMD contains the interactive map exploration functions. Examples of timelines are shown elsewhere in this report in Figures 36, 37, and 66. These functions include the ability to identify groups of papers on the timeline by drawing a box around

them with a mouse. which allows listing the papers in a selected group, listing a frequency table of selected references in a group, and drawing links between selected papers on the map.

Table 10. List of mapping routines used in the software.

ROUTINE	DESCRIPTION
BIB_LABELS	Put bib coupling cluster labels on a current figure
CLEAN_AR_NAMES	Disambiguate reference author names
COMBINE_ALIAS	Disambiguate p_ar matrix
CROSSHAIR	Put crosshairs on a crossmap
DELETEMAP	Sets deleted map pointer to 0
FIG2HTML	Converts figure coordinates to html image coordinates
FREQMOD	Make map circles proportional to cited frequency
GCMAP	Get the current tf map
GETBOX	Gets a box on current fig
HISTORIOGRAM	Puts connection matrix of historiogram in project
MAKE_TIMELINE_HTML	Make timeline webpage
MAKE_XMAP_P_AP_HTML	Make html crossmap for paper to paper author
MAKE_XMAP_P_AR_HTML	Make html crossmap for paper to ref author
MAKE_XMAP_P_R_HTML	Make html crossmap for paper to references
MAKE_XMAP_YP_AR_HTML	Make html reference author usage map
MAKE_XMAP_YP_R_HTML	Make html usage map for references
MAKEDENDRO	Makes the set of lines for a dendrogram
MAP_CMD	Function which draws a map
ORDERDENDRO	Find order for a dendrogram plot.
RECLUSTER	Apply a new threshold to hierarchically clustered data
REDSIM	Reduces a similarity matrix
STUBZ	Function to convert tree to a shorter tree
SUBTREES	Finds the start and ending leaves of dendrogram subtree
TMAP	Make timeline plot
TRANSZ	Translate output of LINKAGE into another format.
TSPORDER	Seriate a dendrogram using simulated annealing algorithm
XMAP	Plot frame crossmap
XMAPDOT_ASSIGNEE	Make patent to assignee crossmap
XMAPDOT_ID	Make paper to ID terms crossmap
XMAPDOT_ONEWORD	Make paper to one word terms crossmap
XMAPDOT_P_AP	Make crossmap paper to paper author
XMAPDOT_P_AR	Make crossmap paper to ref author
XMAPDOT_REF	Make paper to reference crossmap
XMAPDOT_REF_PAT	Make a patent to ref patent xmap
XMAPDOT_YP_AR	Make ref author usage plot
XMAPDOT_YP_R	Make reference usage xmap
XMAPDOT_YR_R	Make reference year timeline

Most of the other routines in this group are concerned with the construction of crossmaps. The crossmap routines work with clusters from a pair of entity-types. A correspondence function is used to measure the relation of clusters of one entity-type to clusters of the other entity-type (Morris & Yen, 2004). The MAKEDENDRO routine is used to place dendrograms on the x and y axis of crossmaps. Crossmaps of different types are very similar in appearance, examples of several types are shown in this report as noted in the list to follow. The following crossmaps can be implemented:

- Papers to references (See Figure 40 and Figure 67)
- Paper to paper author (See Figure 42)
- Paper to reference author (See Figure 41 and Figure 69)
- Paper to index term (See Figure 72)
- Patent to patent assignee (not shown)
- Paper to one-word abstract terms (not shown)
- Patent to reference patent (not shown)

In addition, the crossmapping routines are used to produce usage plots of reference usage and reference author usage as described in Section 12.2 and demonstrated in Figures 38, 39, 68, 70, and 71.

Another important set of routines in this group are concerned with producing web page graphics that can be posted and explored by subject matter experts. The most important of these routines is MAKE_TIMELINE_HTML, which builds a timeline webpage and adds links to make database calls to get summary information about specific clusters of papers. Figure 51 shows a timeline as posted on a webpage. The papers are ranked by the number of citations that their corresponding references receive and the ranking of the top 20 papers is placed next to their corresponding symbols on the map. When the user clicks on these numbers, a hyperlink is invoked which executes an ASP program on the server that retrieves summary data about the paper from the database and displays that data on a separate web page for the user. Additionally, the ranked papers are listed in rank order below the timeline map. In Figure 51, the first four listed papers are visible below the map. A very useful feature of this web-based map is the hot links on the right of the map that provide summaries of paper clusters in a separate browser window. Note in Figure 51 that for every horizontal track corresponding to a paper cluster there are 6 hotlinks labeled 'P', 'R', 'AP', 'AR', 'JP', and 'JR.' Clicking on one of these links will execute an ASP program on the server which will execute a query to the database that produces a summary table in a separate browser window:

- **P**: produces a list of papers in the cluster.
- **R**: produces a table of references cited in the cluster, ranked by the number of citations received.
- **AP**: produces a table of paper authors of papers in the clusters, ranked by the number of papers authored.
- **AR**: produces a table of reference authors in the cluster, ranked by the number of citations received.
- **JP**: produces a table of paper journals in the cluster, ranked by the number of papers the journal has in the cluster.
- **JR**: produces a table of reference journals in the cluster, ranked by the number of citation received.

This feature is quite useful for making the results of analysis available to subject matter experts, and allows those experts to explore the collection to label the clusters by topic, and identify important entities in the collection. Web page versions of crossmaps can also be produced.

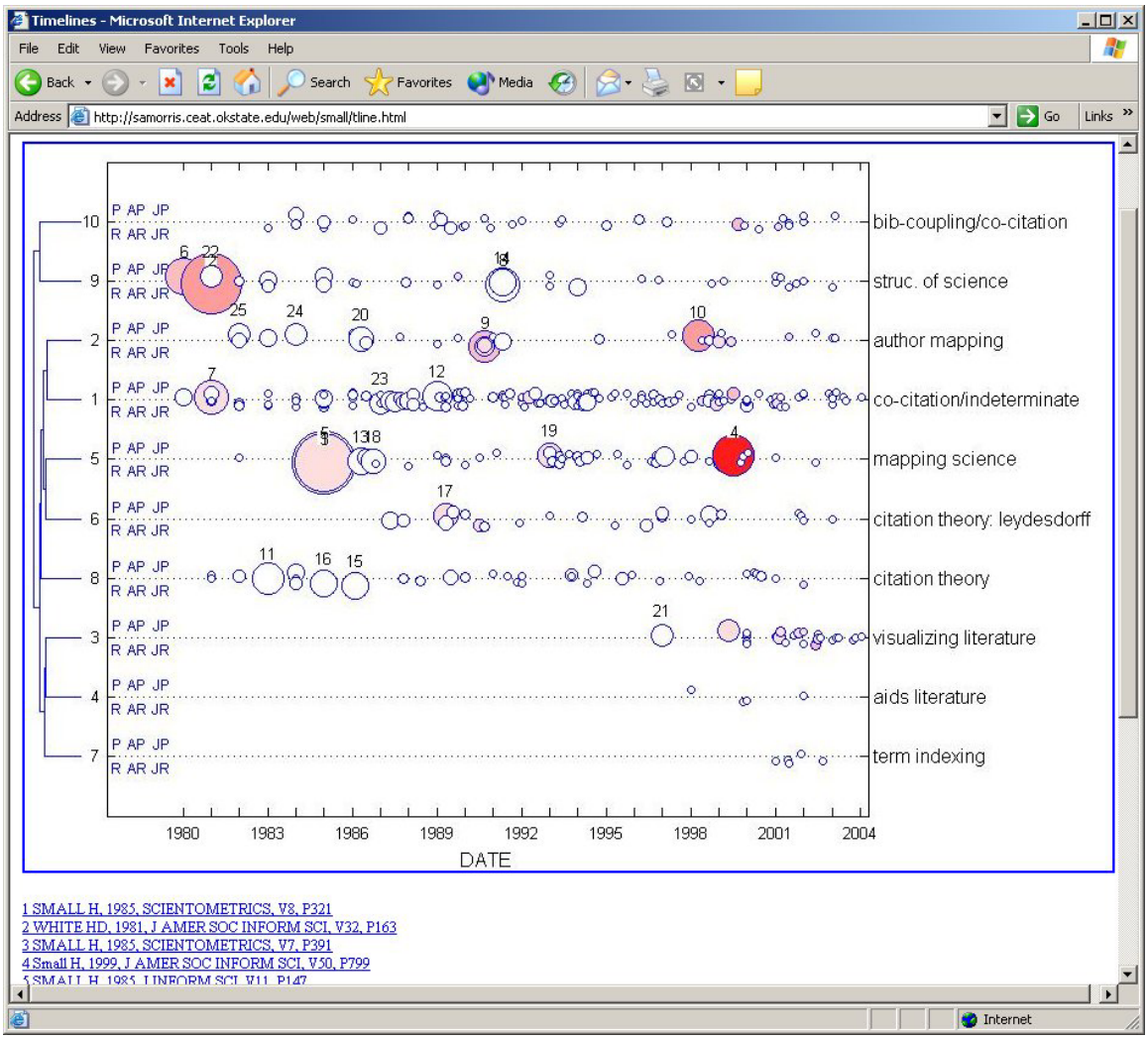


Figure 51. A web-based implementation of a timeline.

14.8 Plotting routines

Plotting routines, listed in Table 11. are used to plot network metrics from the collection of papers. These fall into three categories, dyadic distributions, co-occurrence distributions, and clustering coefficient distributions.

For dyadic distributions, CITES_PER_PAPER plots the reference per paper distribution, while REF_DIST plots the paper per reference distribution of the collection. These two routines are easily adapted to plotting

other dyadic distributions in the collection, but such routines were never added because they are seldom needed. Note that one very useful routine is MAKE_CITE_DIST, a routine that performs maximum likelihood expectation estimation of a zeta (power-law) distribution on a table of frequencies. This routine uses a table described by Goldstein, Morris & Yen (2004) to find the MLE estimate and plots the paper per reference frequencies along with the fitted zeta distribution. Additionally, expected 5% and 95% percentiles are plotted to give users an idea of the expected scatter of points in the plot. Figure 52 shows an example of a plot of a zeta distribution fit to paper per reference frequencies from a collection of papers.

Table 11. List of routines for plotting distributions.

ROUTINE	DESCRIPTION
AUTHOR_PAPER_CLCOF_DIST	Plot co-author clustering coefficient distribution
BIB_COOC_DIST	Plot the bibliographic coupling distribution
CITES_PER_PAPER	Plot reference per paper distribution
CLUST_COEFF	Computes clustering coefficient distribution
HIGHLIGHT	Plot group dots on all dotmaps
MAKE_CITE_DIST	Routine to estimate a zeta distribution
PAPER_REF_CLCOF_DIST	Plot bib coupling clustering coeff distribution
PAPER_REF_COOC_DIST	Plot bib coupling distribution
PLOT_PR	Plot a diagram of the paper reference matrix
PLOT_RANK_AP	Plots map of paper authors, y as dendrogram, x as log of rank
PLOT_RANK_AR	Plot reference authors, y as dendrogram, x as log of rank
REF_COOC_DIST	Plot co-citation distribution
REF_DIST	Plot paper per reference distribution
REF_DIST_COMP	Comparison plot of two paper per reference distributions
REF_DIST_COMP_CUM	Plot comparison of two cumulative paper per reference distributions

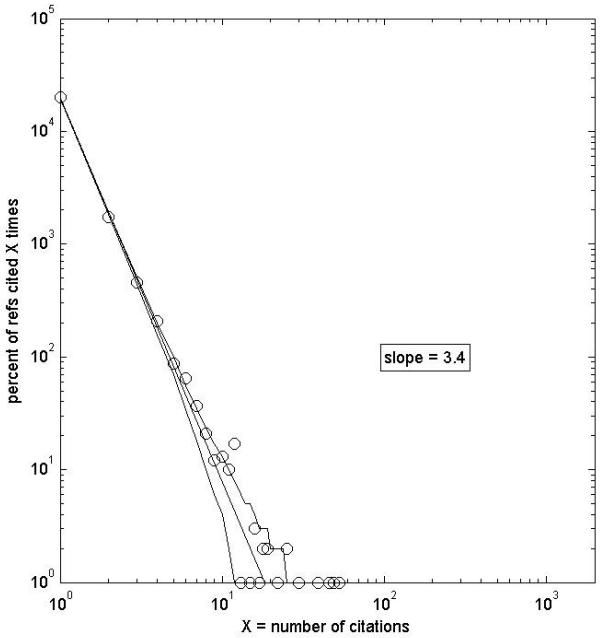


Figure 52. Example of plot of MLE fit of papers per reference using MAKE_CITE_DIST routine.

Several routines are available to plot co-occurrence distributions. Most important among these are BIB_COOC_DIST, for plotting bibliographic coupling per paper pair distribution, and REF_COOC_DIST, for plotting the co-citation per reference pair distribution. Still other routines are available to plot clustering coefficient distributions. For example, AUTHOR_PAPER_CLCOF_DIST plots the clustering coefficient distribution of paper authors linked by coauthorship, and PAPER_REF_CLCOF_DIST plots the clustering coefficient distribution of papers linked by bibliographic coupling.

14.9 Report routines

Routines for producing reports, shown in Table 12, are focused on producing summary information of clusters of entities within the collection of papers. Two types of reports have been found to be useful. These are summary reports of research fronts, that is, papers clustered by bibliographic coupling, and summary reports of co-citation clusters, that is, references clustered by co-citation. Routines that produce similar reports for clustered entities in collections of patents have also been written.

Table 12. List of routines for producing reports.

ROUTINE	DESCRIPTION
DBASE_SELECT	Call dialog box to select current database
DEBLANK	Remove trailing blanks.
PRBIBV2	Prints a report on bib coupling clusters
PRCOCV1	Print a report on co-citation clusters
PRINTABSTRACTSV3	Print bib coupling cluster report with abstracts
PRINTPAT	Print patent bib coupling cluster report
PRPAT_REF	Print patent co-citation cluster report

Table 13 shows an example of a report for research fronts. This is the summary report of a cluster labeled Cluster 7. It starts with a list of the papers in the cluster. This is followed by summary tables: 1) references, 2) paper authors, 3) reference authors, 4) paper journals, and 5) index terms. The report, or sections of this report can be printed out to allow experts to manually browse clusters in the collection.

Table 14 shows an example of a report on co-citation clusters, showing the information presented on one co-citation cluster. This information consists of a list of the references in the clusters. Also, for each research front, the number of citations to references in the cluster is reported along with the ratio of citations to references divided by the number of papers in the research front. This information helps the user to label the co-citation cluster by topic by associating the cluster with the label of the research front whose papers cite it the most. This technique was used by Chen and Morris (2003) to label co-citation clusters on Pathfinder maps of a collection of papers on the subject of botulinum toxin.

Table 13. Example of a report on a research front.

```
--------------------- CLUSTER 7 SUMMARY---------------------


Cluster 7, 8 papers

The endless gallery: Visualizing authors'' citation images in the humanities.   White, H; Lin, X; Buzydlowski, J; P ASIST ANNU MEET,  , vol 38, is null, page 182, 2001

Mining a Web Citation Database for author co-citation analysis.    He, YL; Hui, SC; INFORM PROCESS MANAGE,  , vol 38, is 4, page 491, 2002

Fitting the jigsaw of citation: Information visualization in domain analysis.   Chen, CM; Paul, RJ; O''Keefe, B; J AM SOC INF SCI TECHNOL,  , vol 52, is 4, page 315, 2001

Bibliometric Information Retrieval System (BIRS): A Web search interface utilizing bibliometric research results.   Ding, Y; Chowdhury, GG; Foo, S; Qian, WZ; J AMER SOC INFORM SCI,  , vol 51, is 13, page 1190, 2000

A new technique for building maps of large scientific domains based on the cocitation of classes and categories.    Moya-Anegon, F; Vargas-Quesada, B; Herrero-Solana, V; Chinchilla-Rodriguez, Z; Corera-Alvarez, E; Munoz-Fernandez, FJ; SCIENTOMETRICS,  , vol 61, is 1, page 129, 2004

User-controlled mapping of significant literatures.   White, HD; Lin, X; Buzydlowski, JW; Chen, CM; PROC NAT ACAD SCI USA,  , vol 101, is null, page 5297, 2004

Term co-occurrence analysis as an interface for digital libraries.   Buzydlowski, JW; White, HD; Lin, X; LECT NOTE COMPUT SCI,  , vol 2539, is null, page 133, 2002

Information visualization, human-computer interaction, and cognitive psychology: Domain visualizations.   Boyack, KW; Wylie, BN; Davidson, GS; LECT NOTE COMPUT SCI,  , vol 2539, is null, page 145, 2002


freq    %       REFERENCE
-----------------------------------------------
   7     87      CHEN CM, 1999, INFORM PROCESS MANAG, V35, P401
   6     75      WHITE HD, 1998, J AM SOC INFORM SCI, V49, P327
   5     62      WHITE HD, 1997, ANNU REV INFORM SCI, V32, P99
   4     50      SMALL H, 1999, J AM SOC INFORM SCI, V50, P799
   3     37      MCCAIN KW, 1990, J AM SOC INFORM SCI, V41, P433
   3     37      BRAAM RR, 1991, J AM SOC INFORM SCI, V42, P233
   3     37      KAMADA T, 1989, INFORM PROCESS LETT, V31, P7
   3     37      LIN X, 1997, J AM SOC INFORM SCI, V48, P40
   3     37      NOYONS ECM, 1999, J AM SOC INFORM SCI, V50, P115
   3     37      WHITE HD, 1990, SCHOLARLY COMMUNICAT, P84
   3     37      WHITE HD, 1981, J AM SOC INFORM SCI, V32, P163
   3     37      SMALL H, 1973, J AM SOC INFORM SCI, V24, P265

FREQ    %       PAPER AUTHOR
-----------------------------------------------
   3     37      Lin, X
   2     25      White, HD
   2     25      Chen, CM
   2     25      Buzydlowski, JW

FREQ    %       REFERENCE AUTHOR
-----------------------------------------------
  28    350      WHITE HD
  22    275      SMALL H
  15    187      CHEN CM
  10    125      DING Y
  10    125      GARFIELD E
   9    112      CHEN HC
   8    100      LIN X
   7     87      MCCAIN KW
   6     75      CHEN C
   4     50      FOWLER RH
   4     50      BUZYDLOWSKI JW
```

```
   4     50      BRAAM RR
   4     50      NOYONS ECM
   3     37      HEARST MA
   3     37      BOYACK KW
   3     37      KAMADA T
   3     37      NOWELL LT
   3     37      WISE JA
   3     37      BORNER K
   3     37      SALTON G
   3     37      SCHVANEVELDT RW
   3     37      KOHONEN T


FREQ    %       PAPER JOURNAL
-----------------------------------------------
   2     25      LECT NOTE COMPUT SCI
   1     12      SCIENTOMETRICS
   1     12      PROC NAT ACAD SCI USA
   1     12      P ASIST ANNU MEET
   1     12      J AMER SOC INFORM SCI
   1     12      J AM SOC INF SCI TECHNOL
   1     12      INFORM PROCESS MANAGE


FREQ    %       ID phrase
-----------------------------------------------
   5     62      DIGITAL LIBRARIES
   4     50      NETWORKS
   3     37      SCIENCE
   3     37      INTELLECTUAL STRUCTURE
   2     25      PATHFINDER NETWORKS
   2     25      INTERNET
   2     25      INFORMATION-RETRIEVAL
   2     25      CITATION
   2     25      AUTHOR COCITATION
```

Table 14. Example of a report on a co-citation cluster.

```
---------------------CO-CITATION CLUSTER 3 SUMMARY---------------------

REFERENCE
-----------------------------------------------
  BORGMAN CL, 2000, GUTENBERG GLOBAL INF
  COVI LM, 1999, INFORM PROCESS MANAG, V35, P293
  WHITE HD, 1998, J AM SOC INFORM SCI, V49, P327
  WHITE HD, 1997, ANNU REV INFORM SCI, V32, P99
  BORGMAN CL, 1996, SOCIAL ASPECTS DIGIT
  GINSPARG P, 1994, COMPUTATION PHYSICS, V8, P390
  GARFIELD E, 1979, CITATION INDEXING



FREQ    Bib. Cluster
-----------------------------------------------
  32      1
  14      9
  11      7
   7      3
   1      6
   1      2
```

The software toolkit covered in this chapter has evolved over a period of four years and served as the test bed for many experiments that led to the concepts that form the core of research reported here. It is quite adaptable to new ideas because its implementation in MATLAB allows new ideas to be tested with

minimum programming effort. In addition, the sparse matrix routines of MATLAB make storage and execution efficient, allowing large paper collections to be analyzed.

Other collections of routines have been developed for other research efforts concerning collections of papers. A collection of routines was written to produce pathfinder maps of references, label them using co-citation clustering analysis, and post interactive pathfinder maps as webpages. Another collection of routines allows analysis and production of summary data from many collections of papers simultaneously. These routines have uncovered patterns of distributions within collections of papers that were previously not reported in the literature. For example, the discovery that the reference per paper distribution tends to be a lognormal distribution was exploited by Morris (2004) to build a comprehensive model of the manifestation of the birth of emerging specialties in journal literature . Additionally, the discovery that the paper authors per paper distribution tends to follow a 1-shifted Poisson distribution was exploited by Goldstein, Morris and Yen (in print) to build a comprehensive model of the manifestation of research work teams in journal literature.

15. CASE STUDY: ANTHRAX RESEARCH

15.1 Introduction

The objective of presenting this case study is to illustrate the application of the concepts introduced in this report and to show how those concepts allow the extraction of useful information about research in a specialty from the structure of the network of entities in its journal literature. Chapter 2 of this report showed that a collection of journal papers can be modeled as entities of different entity-types linked in a cascaded bipartite network structure.

The pattern of these linkages are a manifestation of research processes in the specialty. These processes include the social structure of the researchers in the specialty, topics of research, collaboration groups, research paradigms, exemplars, experts, and schools of thought. It is the purpose of this chapter to show that patterns of links can be extracted and the underlying research processes that produced them can be detected with some reliability.

The example presented here is on the topic of anthrax research and covers a period of about 60 years. An initial study on anthrax research was used by Morris, Yen, Wu and Asnake (2003) to show the use of bibliographic coupling to form research fronts of papers. Research fronts were defined as groups of papers that tend to cite common references. Such groups of papers tend to cover a common research sub-topic in the specialty. Morris, et al, showed that timelines of research fronts can be used to visualize structure and dynamic changes in a research specialty. The collection of papers on anthrax studied by Morris, et al, was updated and the analysis of that collection is presented here.

15.2 Background of anthrax research

This section contains an update of the background summary on anthrax research presented in Morris, et al (2003). Anthrax research makes an excellent benchmark for testing the ability to visualize temporal changes in research fronts as they appear in the scientific literature. A great deal of anthrax research has been performed in the past 20 years; it is well documented, and is well covered by the Science Citation Index. A review paper exists (Bhatnagar & Batra, 2001) that names and discusses many key papers in anthrax research in the past 20 years. Modern anthrax research begins in 1946, when anthrax protective antigen was discovered. Early research ranges from 1946 to about 1975 and covers toxin research,

vaccines, inhalational anthrax and medical treatment. After a hiatus of research in the 1970's several research fronts emerged with varied growth characteristics. Vaccine and gene sequencing research fronts have proceeded steadily for 15 to 20 years, for example, while research on anthrax toxins shows a pattern of rapid growth and specialization. The topic of anthrax bioterrorism emerged since 1999 in response to perceived threats, and the Fall, 2001 bioterror attacks through the U. S. postal services have produced a shock to the specialty that generated great interest in anthrax research and produced new research fronts dealing with aspects of anthrax bioterrorism.

The next three paragraphs give a short summary of anthrax and anthrax research which will aid the readers in understanding information presented in this case study. Most of this summary is derived from Bhatnagar and Batra (2001). Anthrax research has a very long and significant history, and was the disease used by Koch, a contemporary of Pasteur in the late 19th century, to prove the original "germ theory." Anthrax was originally thought to cause death by blocking of capillaries, but experiments by Smith and Keppie in the 1950's showed that it kills through the actions of a toxin. The anthrax toxin consists of three parts, *protective antigen* (PA), *lethal factor* (LF), and *edema factor* (EF). Gladstone reported on anthrax protective antigen for the first time in 1946, Smith and Keppie reported that anthrax kills with a toxin in 1954, Beall reported that anthrax uses a three part toxin in 1962, while Leppla reported in detail on lethal factor and edema factor in a seminal paper in 1982. A seminal paper on the efficacy of a human anthrax vaccine was published by Brachman in 1962 and the vaccine itself became available in 1970.

Because of extensive vaccination of animals, and safer handling methods in factories and mills processing animal products, human anthrax became very rare by the 1970's. Also in 1972, the Biological Weapons Convention, endorsed by over 140 nations, prohibited the development of biological warfare agents such as anthrax. Anthrax research slowed down considerably through the 1970's. However, in 1979, a large human anthrax epidemic occurred in Sverdlovsk, a city in the Soviet Union, as a result of the accidental release of anthrax spores from a military biological facility. This event propelled the funding of a new wave of anthrax research starting in the 1980's (Turnbull, 1991). This later research has produced a great deal of progress and has resulted in the development of several new sub-specialties in anthrax research dealing with toxin research, anthrax genetics, anthrax detection, anthrax bioterror, vaccines, and more.

As a disease, anthrax spores enter the host and are taken up by macrophages, amoeboid cells that attack foreign matter in the host, and are transported to nearby lymph nodes. Spores are protected from the macrophages by a *capsule*, an external covering. The bacteria germinate and after release from the macrophages the bacilli multiply in the lymph system and eventually enter the blood stream. Friedlander first reported the importance of macrophages in the spread of the infection in a seminal paper in 1986.

In the bloodstream the bacilli secrete the three-part toxin that eventually kills the host. When attacking cells, protective antigen bonds to a receptor protein on the host cell surface where it cleaves to become the

protein PA63 and then forms a portal into the cell through which lethal factor and edema factor pass to do their damage inside the cell. Anthrax treatment usually fails if delayed because, while it is possible to kill off the bacilli with antibiotics, the toxin that was produced before treatment remains to kill off the host.

15.3 Acquisition and storage of data.

The first acquisition of papers on anthrax research was conducted on December 12, 2001, using ISI's Web of Science (WOS) product, and following the following procedure:

- Using a keyword search on the term "anthrax", 821 papers were acquired, limiting the search to papers available from WOS from 1980 and later.
- A frequency table of references was constructed from these 821 papers, and a list of the top 50 cited references was constructed. Any paper corresponding to a reference in this list that was not in the collection was acquired from WOS if available. This brought the collection to 833 papers.

These 833 papers were used for the original study by Morris, et al. The collection was updated on February 25, 2003, using the search term "anthrax OR anthracis" to find all WOS papers available from 1945 forward. When this collection was combined with the previous collection, the anthrax collection totaled 2472 papers. Figure 53 diagrammatically lists the number of entities and links in the collection.

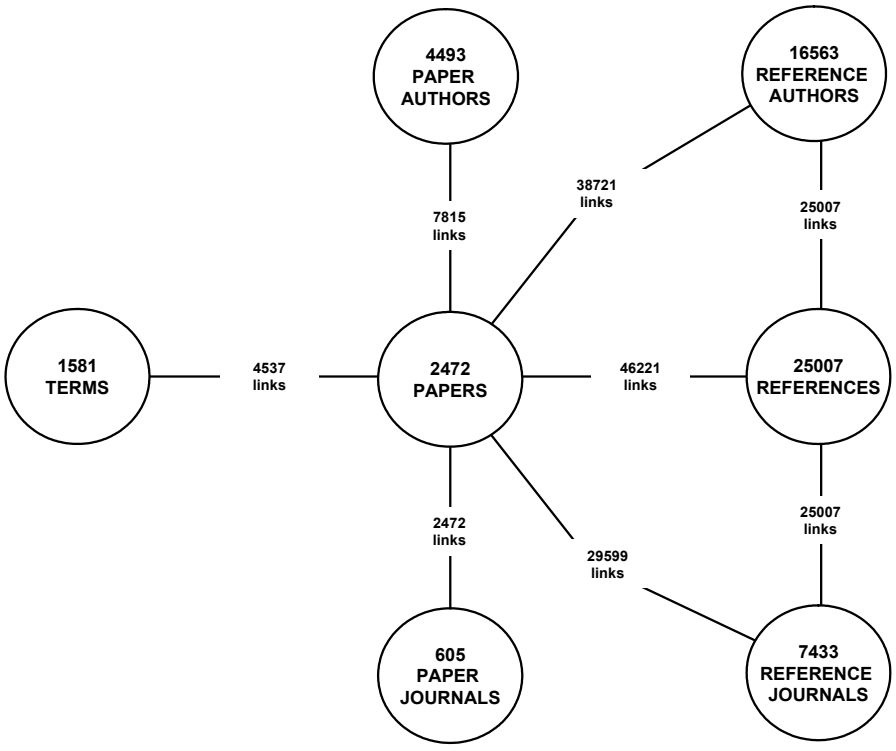


Figure 53. Diagram showing the number of entities and links in the anthrax collection.

The data was loaded into five tables in an MS Access database in the following tables:

- A table for papers and paper attributes.
- An author table whose records correspond to links from paper authors to papers
- A citation table whose records correspond to links from papers to references
- A reference table which holds reference keys, links to reference authors, reference journals and reference year.
- A terms tables holding links between terms and papers.

The data was loaded into MATLAB into the following matrices:

- **O**[*p*;*r*]: paper to reference matrix
- **O**[*p*;*ap*]: paper to paper author matrix
- **O**[*p*;*ar*]: paper to reference author matrix
- **O**[*p*,*t*]: paper to term matrix
- **O**[*p*, *jp*]: paper to paper journal matrix

An initial exploratory analysis of the data was done by clustering papers on bibliographic coupling, using a threshold of 5 common references per paper. Examination of these clusters showed that the papers generally fell into coherent research topics and did not show any off-topic clusters of papers that needed to be discarded. Initial timeline mapping of the papers indicated that the data fell into two distinct periods: 1) early research from 1945 to 1975, and 2) current research from 1976 to the present. Early research fell into three categories: toxin research, medical treatment, and vaccine research. Current research, through increased specialization, includes several topics: toxin research, vaccines, gene sequencing, strain identification, and bioterrorism. Detailed discussion of the final timeline appears in Section 15.5.

15.4 Exploratory data analysis

Figure 54 shows the reference per paper distribution for papers from 1945 to 1975. This approximates a lognormal distribution with a mode of about 10.3 and a mean of 12.7 references per paper. However, as shown in Figure 55, in the period from 1976 to 2003, the mean increases to about 30 references per paper. This increase probably represents the results of four processes: 1) the body of knowledge concerning anthrax increased over time, forcing authors to increase the number of references made in each paper to orient the reader about the position of the paper in the specialty, 2) the social conventions regarding citation of references changed over time; scientists now tend to cite more references than in previous years, 3) the increased specialization of anthrax research caused scientists to use journals whose editorial standards for

citing encourage larger numbers of references per paper, or 4) the journal coverage of the Science Citation Index over the years is biased in later years to journals with higher average references per paper.

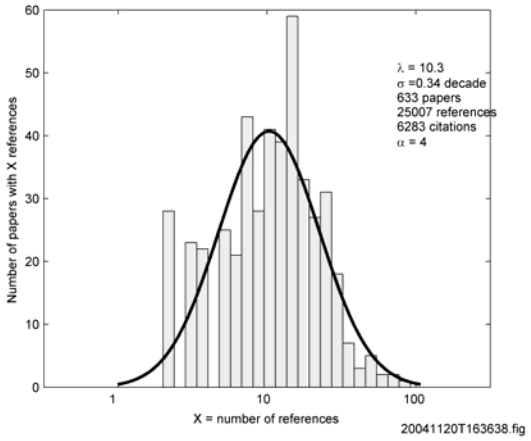


Figure 54. Reference per paper distribution for the period from 1945 to 1975.

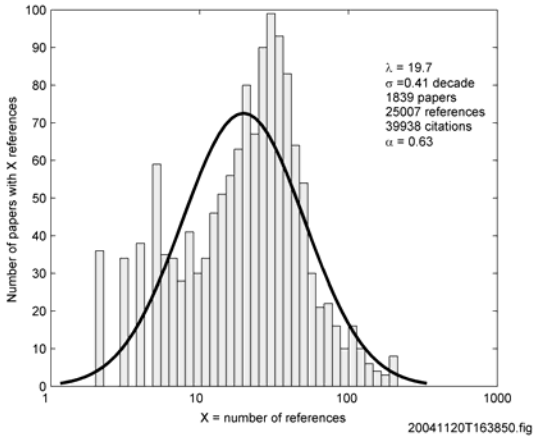


Figure 55. Reference per paper distribution for the period 1976 to 2003.

Figure 56 shows a plot of papers per reference for papers from 1945 to 1975. This distribution approximates a zeta (power-law) distribution with an exponent of 2.65. Figure 57 shows the paper per reference distribution for the period 1976 to 2003. This distribution also approximates a zeta distribution except that there is distortion in the tail showing that there are more heavily cited references than would be predicted by the zeta distribution. This shows that, compared to the earlier period, there are more 'exemplar' references in the collection, representing a greater consensus on the base knowledge in the specialty that authors cite to establish background knowledge for their papers (Hargens, 2000).

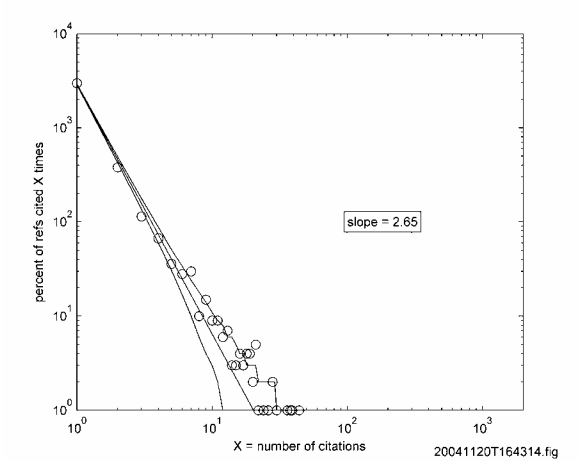


Figure 56. Paper per reference distribution for the period 1945 to 1975.

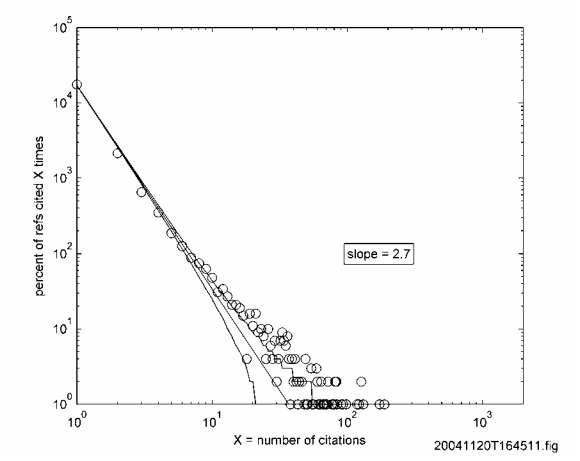


Figure 57. Paper per reference distribution for the period 1976 to 2003.

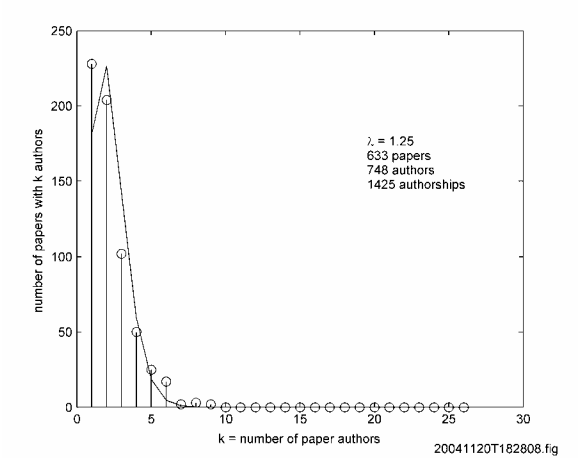


Figure 58. Paper author per paper distribution for the period 1945 to 1976.

Figure 58 shows the paper authors per paper distribution for the period 1945 to 1975. A fitted 1-shifted Poisson distribution is also shown. There is a good general fit of the data to a 1-shifted Poisson distribution. The mean authors per paper is 2.2.

Figure 59 shows the paper author per paper distribution of the period 1976 to 2003. The fitted 1-shifted Poisson distribution is plotted. This distribution indicates an inflated number of single author papers, which indicates that the two processes are driving the author per paper distribution. One process approximates a 1-shifted Poisson distribution while a second process produces only single author papers. Thus, the distribution is analogous to a zero-inflated Poisson distribution (Lambert, 1992). Note that in this period the mean authors per paper has risen to 3.4. This indicates the size of research teams increased over the first period, probably a result of better funding in the second period (Beaver, 1978).

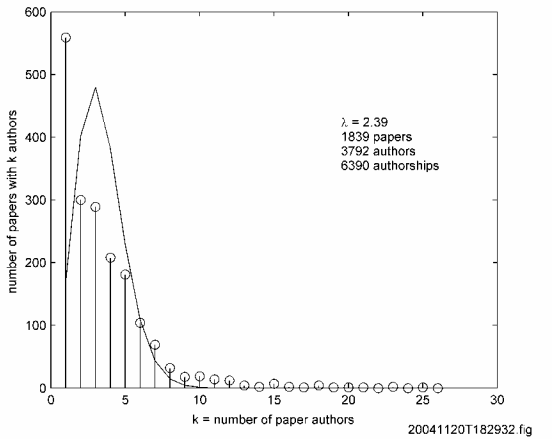


Figure 59. Paper author per paper distribution for the period 1976 to 2003.

Figure 60 shows the paper per author distribution for the period 1945 to 1975. In accordance with Lotka's Law (Lotka, 1926), the distribution follows well a zeta distribution. The estimated zeta distribution exponent is 2.35. This indicates a well defined group of core researchers in the specialty. Figure 61 shows the same plot for the period 1976 to 2003. In this case, the distribution also well approximates a zeta distribution. Note that the fitted exponent is 2.6. This indicates that the later period has a slightly larger set of core researchers than earlier.

Figure 62 shows the number of papers as a function of paper year. Papers per year is fairly constant through the 1950's and 1960's with a lean period in the early 1970's. From the early 1980's, there is slow growth. In 2001 the growth becomes dramatic, doubling the number of papers from 2000 to 2001, and doubling the number yet again in 2002. This exponential growth was in response to dramatic discoveries in toxin research, interest in vaccines and , in 2002, intense interest focused on the bioterrorism postal attacks.

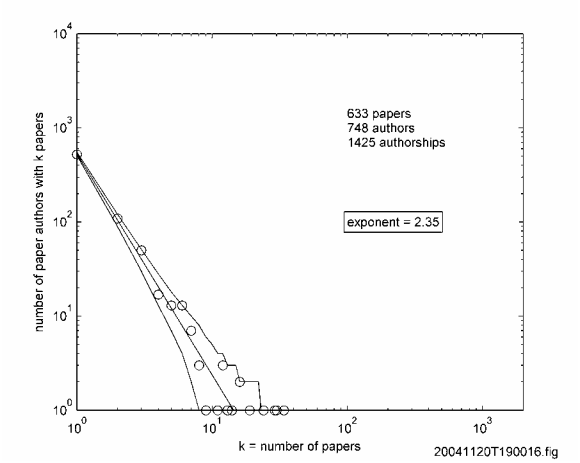


Figure 60. Paper per paper author distribution for the period 1945 to 1975.

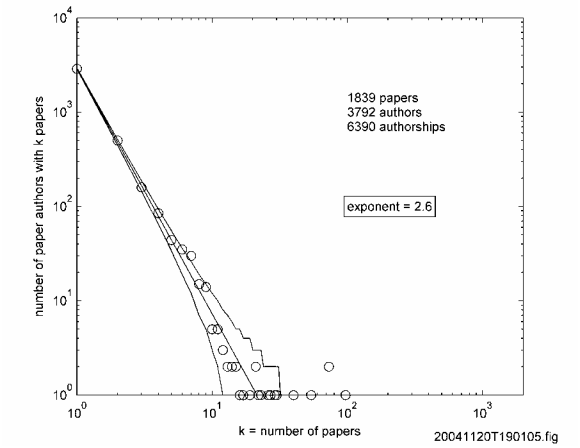


Figure 61. Paper per paper author distribution for the period 1976 to 2003.

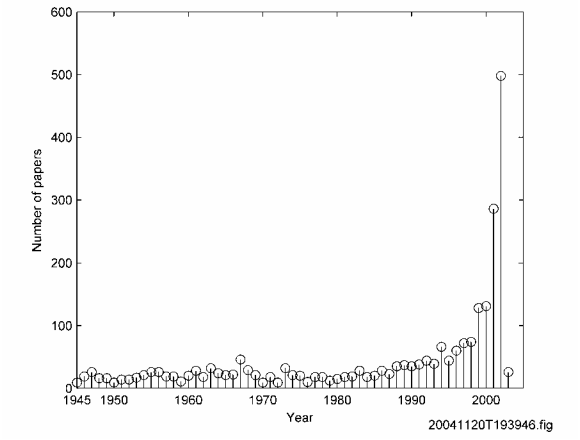


Figure 62. Paper per paper year plot of the anthrax collection.

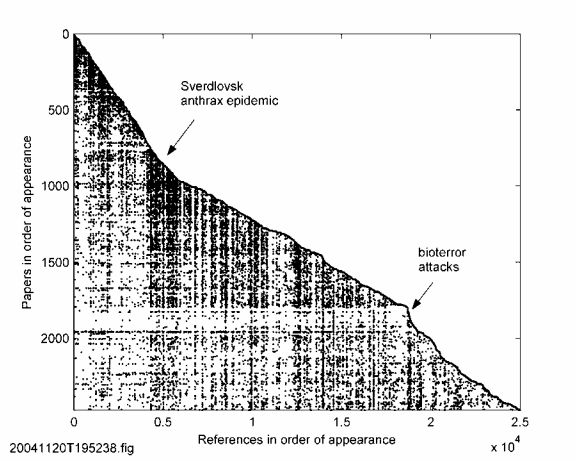


Figure 63. Diagram of the paper to reference matrix for the anthrax collection.

Figure 63 shows a diagram of the paper to reference matrix for this collection. This is a very interesting diagram which shows marked changes in the dynamics of the collection in response to events in the specialty. Initial research from 1945 to 1975 is from paper 1 to paper 800. References appearing during this period are heavily cited, judging from the density of dots in the diagram. A change in the specialty occurs at about paper 750 or so. A new group of references appears and the citations to previous references thins considerably. The rate of appearance of new references accelerates. This change in the specialty was probably in response to large increases in government funding of anthrax research in the late 1980's in response to bioterrorism threats. Another shock to the specialty occurred in the fall of 2001 with the occurrence of the postal bioterror attacks. This is noted in the diagram. This contributes to the great volume of papers published in 2002. After the attacks, a large number of papers appear that do not cite many previous references. This produces an empty horizontal band in the matrix. At the same time, a small number of references appears that become heavily cited, indicating the creation of a new set of exemplar references corresponding to new research fronts. Figure 64 shows a plot of the number of references as a function of the number of papers. In this diagram the beginning of each year is marked as a vertical line. Note that the change in the first initial increase occurs about 1990, while the second event occurs in 2001 and corresponds to the postal attacks.

Finally, Figure 65 shows a plot of the paper per paper journal distribution. which well approximates a zeta distribution with an exponent of 1.9. This is in accordance with Bradford's Law (White & McCain, 1989), which predicts a set of core journals for a specialty.

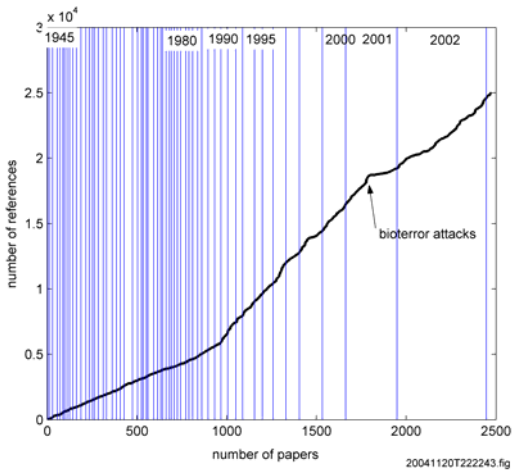


Figure 64. Number of references as a function of the number of papers in the anthrax collection, with vertical lines showing the start of each year.

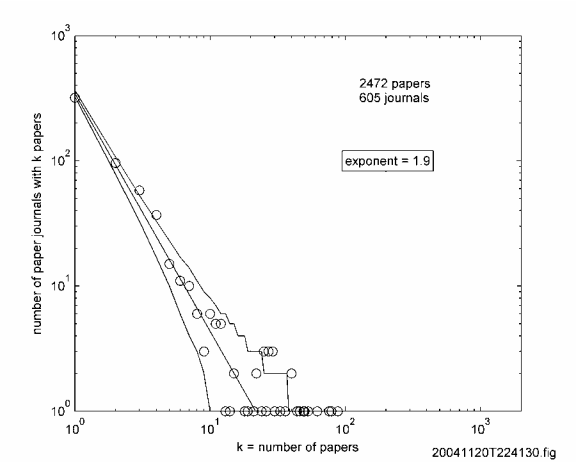


Figure 65. Paper per paper journal distribution for the anthrax collection.

15.5 Research front timeline

The paper to reference matrix consists of 2472 papers having 46221 links to 25007 references. Using a co-occurrence threshold of 5 common references, 987 papers were clustered into 35 research fronts. Figure 66 shows a research front timeline for the anthrax collection. The figure is rotated 90 degrees for a better fit on the page. Looking at the left, the identifying numbers for each research front are printed in a column to the right of the clustering dendrogram. The papers in each research front are plotted by time in horizontal tracks, with the research front labels on the right side of the plot. Research front labels were found by manually searching the papers in each research front for themes. The circles on the plot correspond to papers and the size of each circle is proportional to the number of times that the paper was cited. Each circle is shaded red in proportion to the number of times its corresponding paper has been cited in the last year of the collection (February 2002 to February 2003).

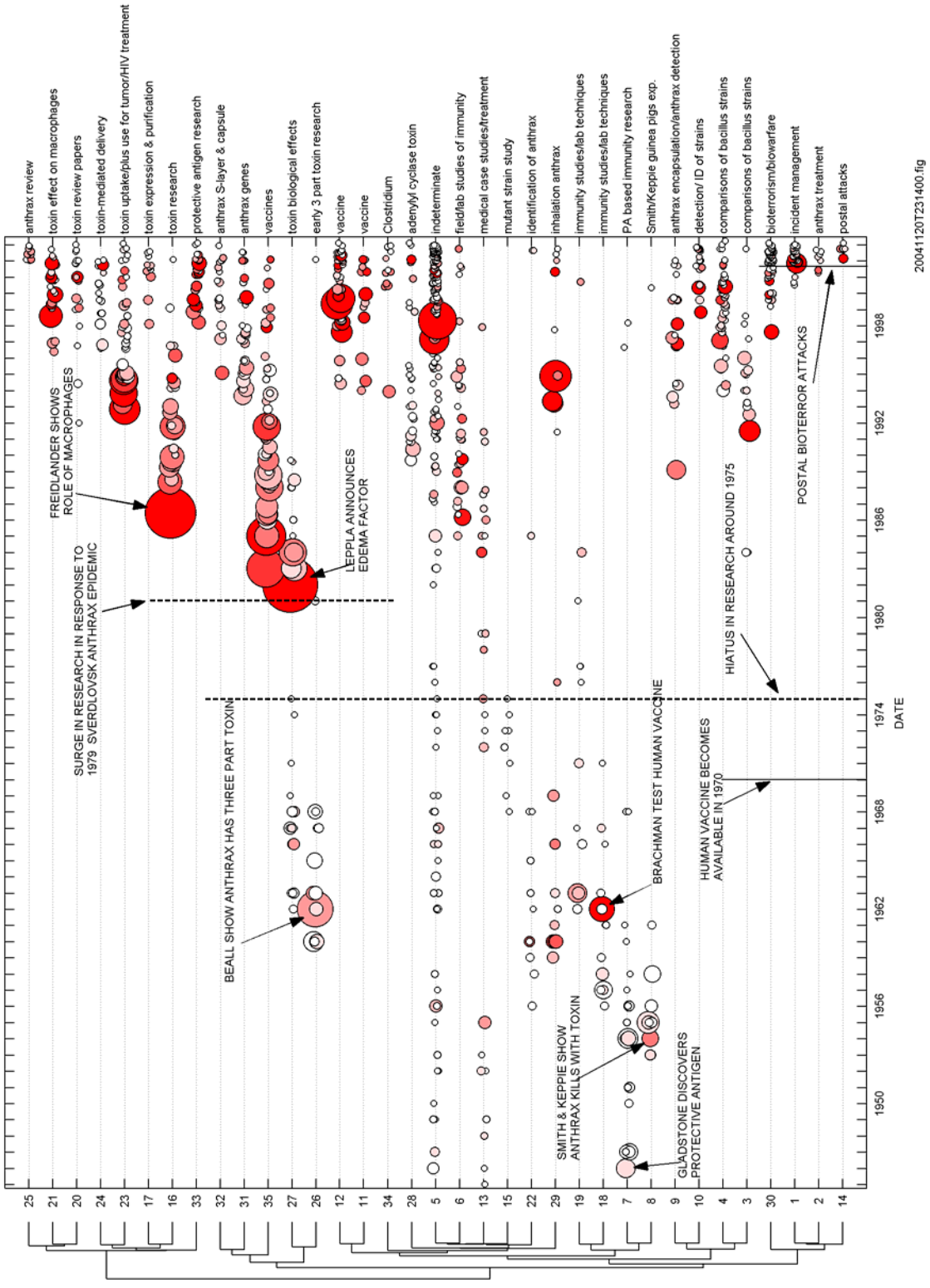


Figure 66. Research front timeline for the anthrax case study.

It is easy to see from the timeline that the collection falls into two distinct sections 1) research from 1945 to about 1975, and 2) research from 1976 to 2003. It seems probable that the great surge in the number of research papers starting in the 1990's is tied to government funding of research in response to bioterrorism threats. The research fronts can be classified as follows:

- Early research (1945 to 1975)
    - Research fronts 7 and 8 are the earliest research and cover early immunity studies and Smith and Keppie's seminal guinea pig experiments that showed that anthrax kills with a toxin.
    - Research fronts 19 and 18 cover vaccine.
    - Research front 29 deals with medical treatment of inhalational anthrax and may be tied to funding of bioweapons research in the 1960's. The work picks up again in the 1990's, an event that may be tied to government funding of research on bioterrorism in the late 1980's. The key papers in this research front are currently being heavily cited in response to the postal bioterror attacks and resulting intense interest in treating inhalational anthrax.
    - Research fronts 15 and 22 are papers that discuss identification of anthrax and discrimination of anthrax strains.
    - Research front 13 consists of papers on medical case studies. This research front continues up to the 1990's, with a dry spell during the late 1950's and 1960's. It is finally superceded around 1985 by research front 6.
    - Research fronts 27 and 26 are papers that continue Smith and Keppie's experiment on the anthrax toxin. These papers establish the knowledge on the 3 part anthrax toxin until they are superceded by the seminal work of Leppla in 1982 and Freidlander in 1988.

- Current research (1976 to 2003)
    - Research fronts 25 to 33 generally cover the topic of anthrax toxin research. These include research on the three parts of the toxin: protective antigen, edema factor, and lethal factor.
    - Research fronts 24 and 23 cover research on using protective antigen to inject materials into cells to induce immunity to AIDS and other diseases.
    - Research fronts 32 to 35 cover general anthrax research in the 1980's, anthrax gene sequencing, and the anthrax capsule.
    - Research fronts 12 and 11 cover current research in anthrax vaccines.
    - Research fronts 34 and 28 are miscellaneous topics. Clostridium is a bacteria that produces botulism toxin and is related to anthrax through bioterrorism research, adenylyl cyclase toxin is closely related to edema factor research.

- Research fronts 9 to 3 cover research on various methods of sensing anthrax and classifying anthrax by strains.
- Research fronts 30 to 14 and 25 are bioterrorism related. Research front 30 covered general bioterror research until the postal terror attacks. The postal attacks induced four other research fronts: 1) Research front 1 on incidence management, 2) Research front 2 on treatment of the disease, 3) Research front 14 covering the postal attacks themselves, and finally 4) a series of reviews on anthrax toxin research which was clustered at the top of the figure into the research fronts covering toxin research.

15.6 Analysis of references

The paper to reference matrix consists of 2472 papers having 46221 links to 25007 references. An occurrence threshold of 40 was used, which yielded 70 highly cited references that were clustered down to single references. Figure 67 shows a crossmap of research fronts to references. References are mapped as columns in the crossmap, with a dendrogram at the top of the figure and reference labels at the bottom of the figure. Research fronts are mapped as rows on the crossmap, with clustering dendrogram on the left and research front labels on the right. Given research front *i*, and reference *j*, the size of the circle on the map at row *i*, and column *j*, is proportional to the percentage of papers in research front *i* that cite reference *j*.

In this map it is easy to see the overlapping correspondence of reference clusters to research fronts:

- At the bottom left of the map note a series of references that are key references for anthrax bioterrorism. This group of references starts with reference number 71 and ends with reference number 34 as seen at the top of the map. Reference 71, Jernigan 2001, is a notable reference that reports on 10 cases of inhalation anthrax from the postal bioterror attack. Inglesby, 1999, reference 3, is a policy paper on anthrax bioterrorism.
- At the far left are a series of four references: Sambrook 1989, Ash 1991, Keim 2000, and Keim 1997, that are key references for methods of detecting anthrax and discriminating among strains.
- At the top of the map a series of 23 references comprise the key references on the topic of toxin research. This group starts with reference 14 on the left and ends with reference 68 on the right. There is a great deal of overlap in the correspondence of these references to different research fronts in toxin research (research front 25 at top down to research front 33.) The key references for all of toxin research are Leppla 1982 and Friedlander 1986, references 15 and 7 respectively. Another important reference is Duesbery 1998, reference 14, corresponding to a paper that explained the mechanism by which anthrax lethal factor kills cells.

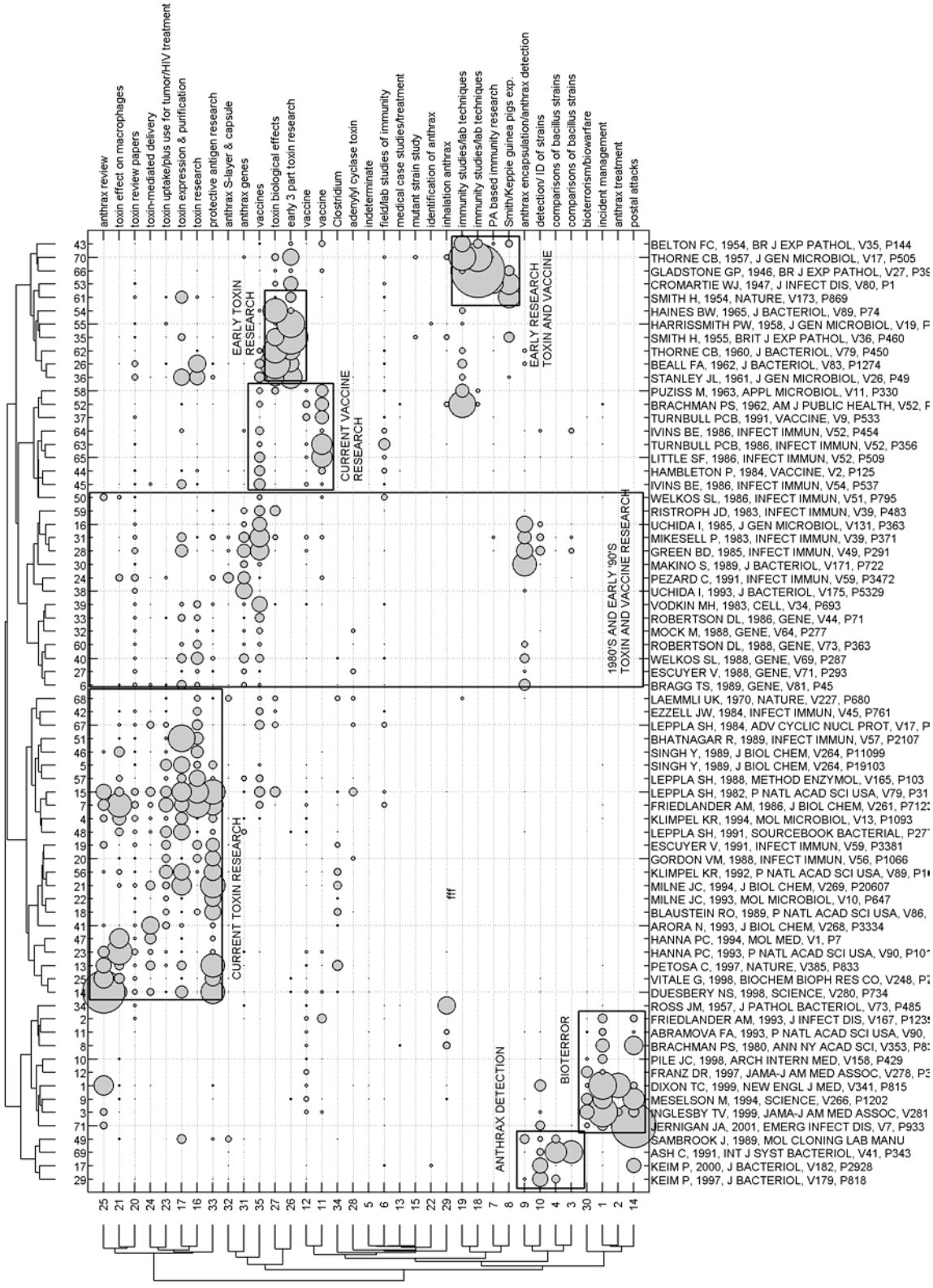


Figure 67. Research front to reference crossmap for the anthrax collection.

- In the center of the map, references from 6 on the left to 50 on the right, correspond to papers published in the 1980's and 1990's in anthrax before a lot of specialization occurred. These are cited from many research fronts, but particularly from research fronts on vaccines and anthrax genetic sequencing.
- Center right on the map a series of references (from 45 on the left to 58 on the right) are key references for current vaccine research. Among these is Brachman 1962, which is reference 52, which corresponds to a report on efficacy of the anthrax vaccine that is commonly used today.
- On the right of the plot are references that were used by papers in research fronts for early anthrax research from 1945 to 1975. The group of 5 references at the extreme right corresponds to the earliest research, and includes Smith and Keppie's original study that showed that anthrax kills with a toxin, Smith 1954, reference 61. Immediately to the left of this group are references, from reference 36 on the left to reference 54 on the right, used by papers in the research front from the 1950's that established that the anthrax toxin has three parts.

Figure 68 shows a map of reference usage for the anthrax collection. In this plot the references arrayed on the x axis are identical to those from the research front to reference crossmap, Figure 67. In this plot the rows correspond to paper years and the columns correspond to references. Given year *i* and reference *j*, the size of a circle at row *i* and column *j* on the map is proportional to the number of times that reference *j* was cited in year *i*. The main purpose of this map is to show obsolescence of references. Because of the small volume of papers in the early research period from 1946 to 1975, the size of the circles in this period are magnified 4 times over the sizes in the later period. The following features are visible on this map:

- on the extreme right the series of references from 62 on the left to 43 on the right have become obsolete. They cease to be cited around 1968, a year which may have corresponded to cuts in funding of anthrax research. These references are not cited much even after anthrax research picks up again in the late 1980's and early 1990's.
- Two references from early research, Stanley 1962 and Beall 1962, references 36 and 26, correspond to papers that characterize anthrax lethal factor toxin. As shown on the plot, these references are still current and being cited to the present day.
- Reference 34, Ross 1957, corresponds to a paper on how inhalational anthrax develops in the lungs. After a long period of no citation that started about 1966, this reference is being cited heavily since the postal bioterror attacks because of the current intense interest on treating inhalational anthrax that resulted from those attacks.

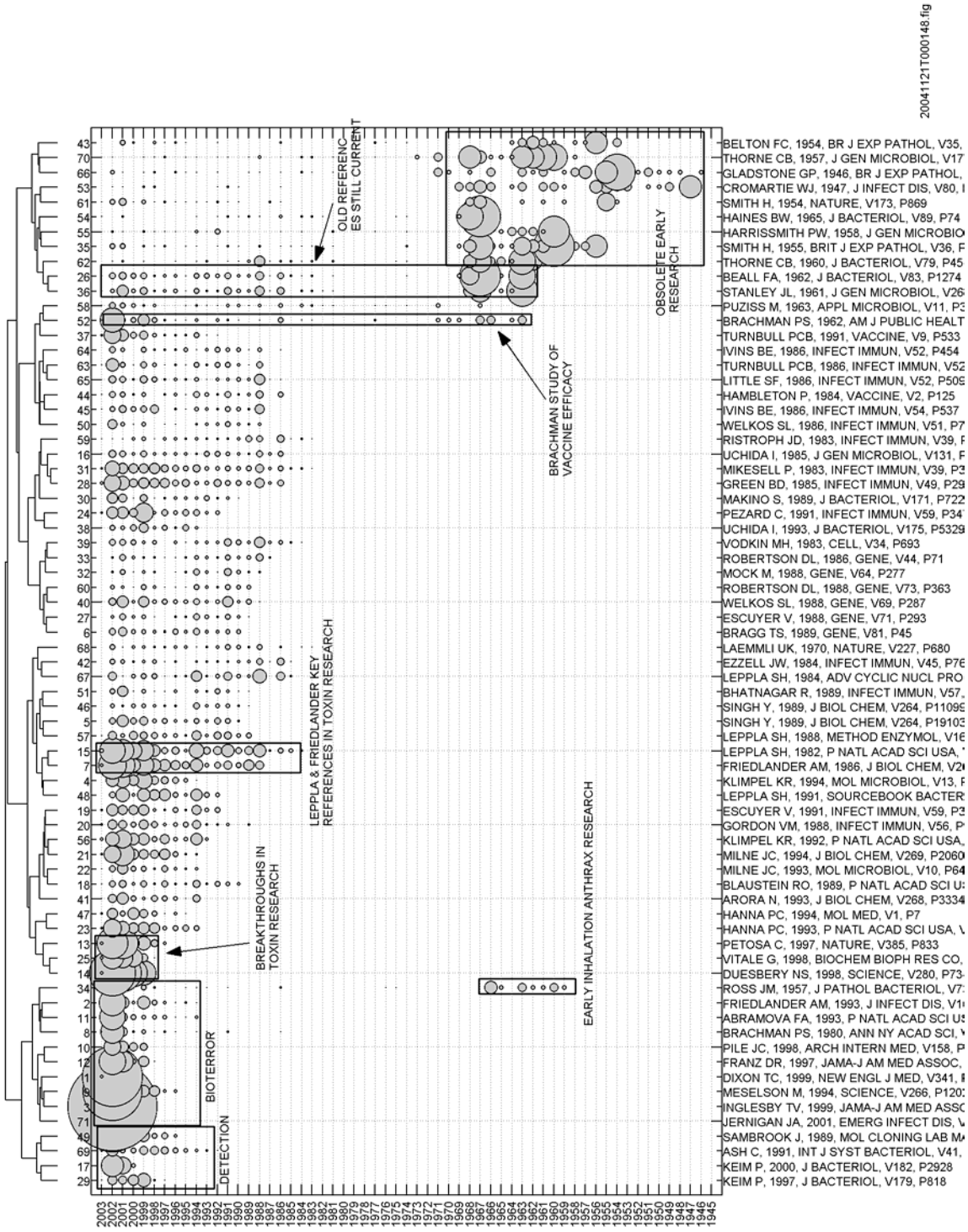


Figure 68. Reference usage plot for the anthrax case study.

- Note that it is easy to see the emergence and use of key references on the detection of anthrax and discrimination of strains of anthrax. These are the references on the extreme left (starting with reference 29 on the left to reference 49 on the right), and the initial reference in this group is Ash 1991 (reference 69).
- Note the two seminal references for toxin research are in the center of the plot. These are Leppla 1982 and Friedlander 1986, reference 7 and 15 respectively. Leppla corresponds to a key paper on edema factor, while Friedlander announces the discovery of the role of macrophages in the spread of the disease within a host. While the Leppla paper appears in 1982, it does not start to be heavily cited until 1988, a 6 year delay. The year 1988 may be a year in which government funding of anthrax research was increased dramatically in response to bioterror threats.
- Finally, note the great number of citations received by key bioterror references in the year 2002. These are to the left of the plot, from reference 71 on the left to reference 34 on the right. This heavy number of citations reflects the intense interest in anthrax bioterror after the postal attacks in late 2001.

15.7 Analysis of reference authors

The reference author matrix consists of 2472 papers having 38721 links to 16563 reference authors. The paper to reference author matrix was treated as unweighted, i.e., all the link weights were set to unity. An occurrence threshold of 40 was used, which yielded 88 highly cited authors that were clustered down to single authors. Figure 69 shows the research front to reference author crossmap. The following features can be seen on this map:

- The reference authors corresponding to early research from 1946 to 1975 have fallen on the extreme left of the plot, in the section with author 68 on the left to author 26 on the right. Within this section of the map, there is a group of authors (from author 78 on the left to author 30 on the right) that are cited from research fronts on immunity research. A second group of authors, from author 60 on the left to author 26 on the right, are cited from the research front on toxin research.
- Moving right, there is a group of authors, from author 58 on the left, to author 71 on the right, that are cited heavily from the research front on anthrax detection.
- Moving right, a group of reference authors from author 77 on the left to author 6 on the right are cited heavily from vaccine research fronts. a subset of these authors from author 2 on the left to author 6 on the right appear to be heavily cited from the most current vaccine research fronts.

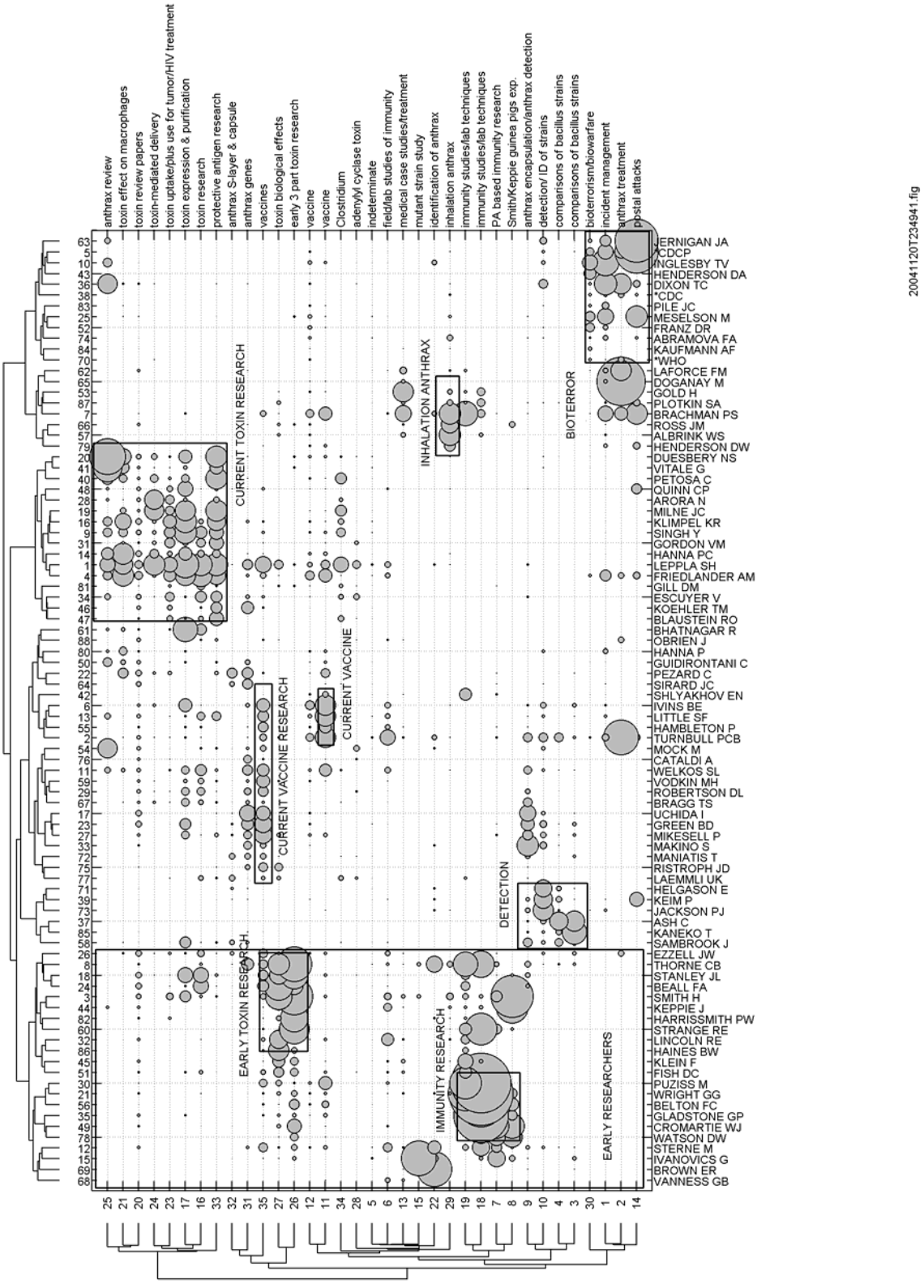
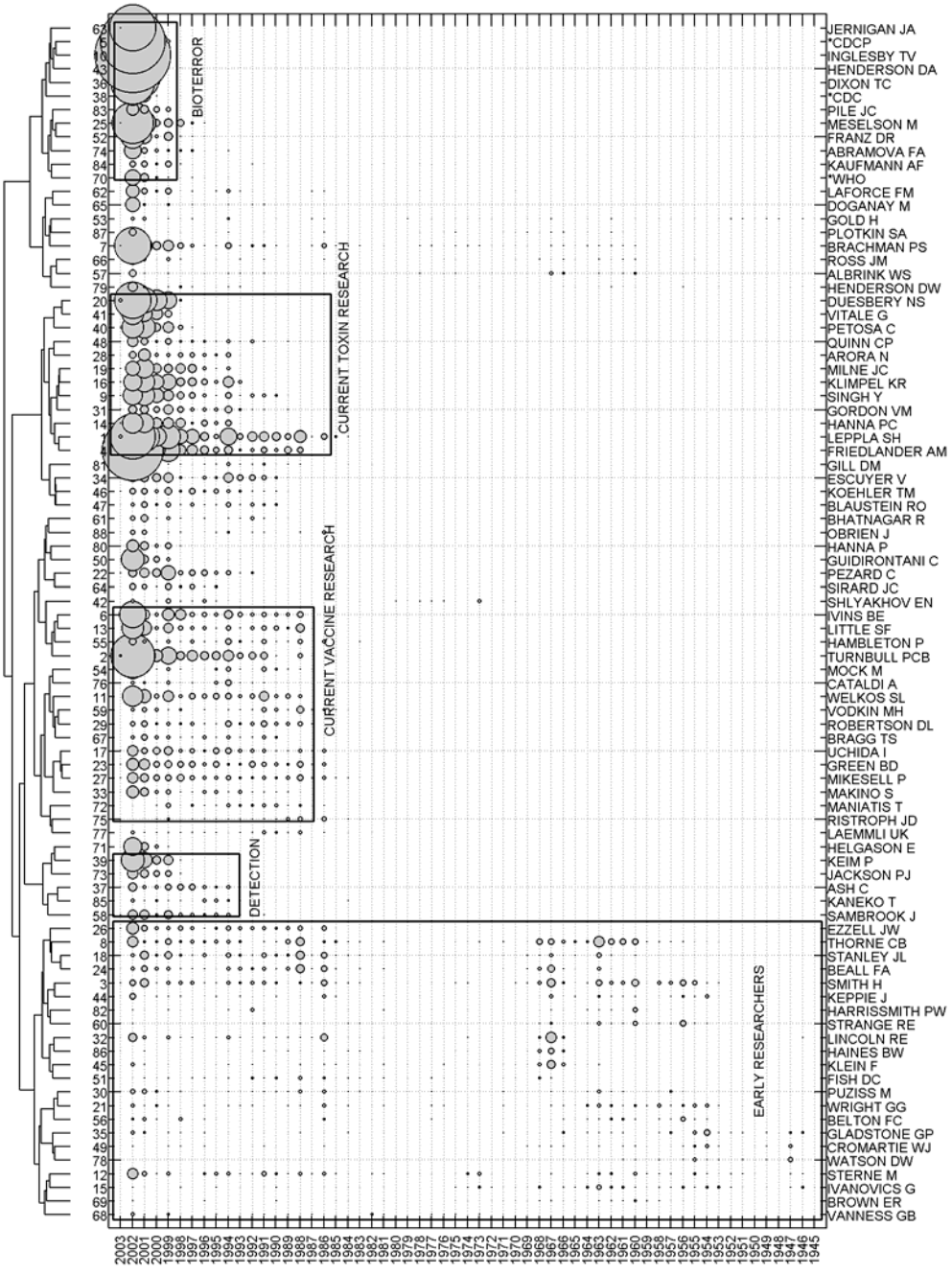


Figure 69. Research front to reference author crossmap for the anthrax collection.



20041121T000427.fig

Figure 70. Reference author usage plot for the anthrax case study.

- Research fronts in toxin research cite a series of authors to the right center in the plot, starting with reference 47 on the left to reference 79 on the right. Note the prominence of both Friedlander and Leppla in this group. Friedlander and Leppla get overlapping use from other research fronts as well, particularly from vaccines and bioterror research fronts.
- Authors representing bioterror topics appear on the extreme right, from author 70 on the left to author 63 on the right. Note the heavy overlap of authors, from 79 on the left to 62 on the right, authors being used by research fronts on both the inhalation anthrax research front but from the bioterrorism research front as well.

Figure 70 shows a usage plot of reference authors which has the following features:

- On the right is the section corresponding to reference authors for early research from 1946 to 1976. Many of these authors remain well cited even in 2002.
- Note that most of the reference authors in current vaccine research and current toxin research begin to be cited in 1988, again suggesting that the volume of papers on anthrax topics increased dramatically as a function of increased funding in that year.
- Bioterror reference authors receive massive references in 2002, showing response of the specialty to the bioterror postal attacks.

15.8 Analysis of paper authors

The paper to paper author matrix has 2472 papers linked to 4493 authors through 7815 authorships. Using an occurrence threshold of 8 yielded 83 authors which were clustered down to single authors. Note however, that because of the large number of papers in each research front, the research front to paper author crossmap did not yield much information and is not shown. In general, this type of crossmap can yield better results on smaller collections that are more homogeneous and do not have such large numbers of authors.

Figure 71 shows a paper author usage plot. In this map it is possible to see which researchers are active and which researchers are no longer working in the specialty. At the top of this plot is a clustering dendrogram for the paper authors. The dendrogram seriation routine used for making this dendrogram tends to put the most distinct clusters to the extreme left and right while placing authors that publish many single authored papers in the center columns. Thus, the most distinct groups of paper authors are easy to distinguish on the left and right on the plot. The following features can be distinguished on this plot:

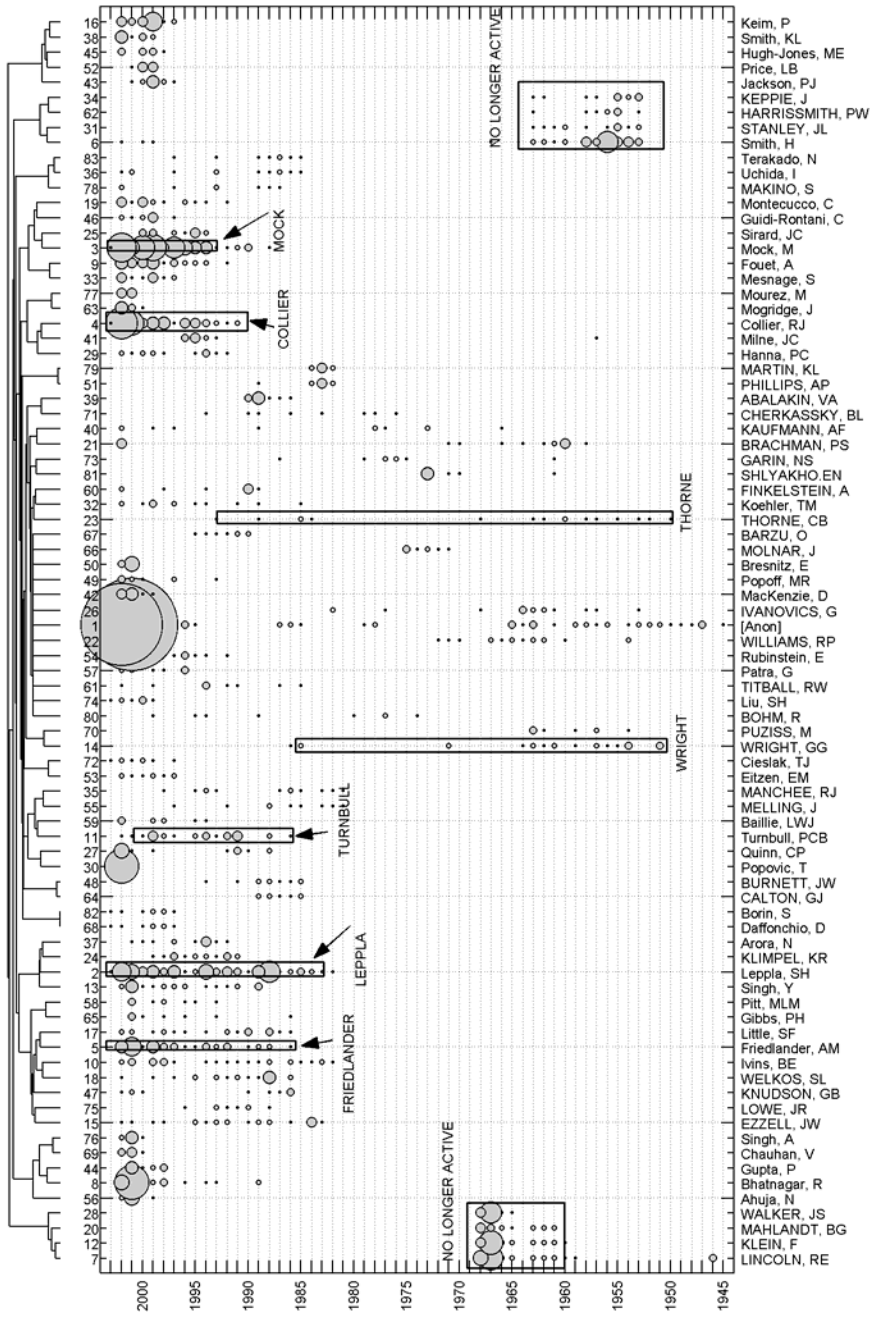


Figure 71. Paper author usage plot for the anthrax case study.

- Two groups of authors can be seen that were active in early research from 1945 to 1975 that are no longer active. On the left, Lincoln, Klein, Mahlandt and Walker form a team which started activity in 1961 and ended activity in 1968, publishing papers on early toxin research and immunology studies. On the right are authors Smith, Stanley, Harrissmith and Keppie, who performed the original toxin research that identified the three part toxin in the 1950's. These authors became active in 1953 and ceased activity in 1963. Several authors in the center of the plot became active in the 1950's and 1960's. Wright (1951 – 1986) and Thorne (1950-1993) had very long publishing careers.
- Note that the author "[Anon]" in the center of the plot is associated with reports from the Center for Disease Control and other government agencies. It would be convenient to add an option to the software to delete such artifacts.
- There are some single authors among the author groups that are very active and have long histories. Friedlander, Leppla, Turnbull, Collier, and Mock are examples noted on the plot. These authors all have publication histories that range back to at least the early 1980's.

15.9 Analysis of terms

The terms used for this study are index terms that are machine-generated terms provided by the Web of Science. These terms are problematic because there appear to be numerous synonyms among them. For example, 'lethal factor' and 'toxin lethal factor' are synonymous but separate terms in this collection. Also, index terms are not provided for papers published before 1991. Despite these problems, the analysis here is shown to demonstrate the usefulness of using terms to assist in validating labels for research fronts.

The paper to term matrix has 2472 papers linked to 1581 terms through 4537 links. The terms per paper distribution (not shown) has a large spike at 10, indicating that the ISI algorithm used to generate terms limits the number of terms per paper to 10. The paper per term distribution (not shown) well approximates a zeta distribution with an exponent of 2.2. Using a co-occurrence threshold of 5 yielded 94 terms, which were clustered down to single terms.

Figure 72 shows a research front to term crossplot for this collection. As in previously discussed crossmaps, a dendrogram is provided on the left of the map for research fronts, with research front labels on the right. The clustering dendrogram for terms is at the top of the map, while the terms themselves are shown at the bottom. Related terms are found by looking at subtrees on the dendrogram. The following features can be seen on this map:

Figure 72. Research front to term crossmap for the anthrax case study.

- a series of terms appears on the far left, term 91 on the left to term 74 on the right, that appear to be terms about toxins produced by various species of bacteria. These terms are associated with the 'clostridium' research front. Clostridium is the bacteria that produces 'botulinum toxin', a term in this group. This group of terms seems to indicate that the clostridium research front may be a group of papers discussing bacterial toxins in general.
- a group of terms, from term 75 on the left to term 72 on the right, appear to be associated with research fronts covering toxin research. However, because many of the terms are specialized, it is difficult to assess the usefulness of the terms in discriminating between topics of research fronts covering toxins.
- a series of terms, from term 76 on the left to term 51 on the right, are terms associated with immunity and vaccines. These terms occur in a research front labeled 'vaccine' and confirm the validity of that label.
- a series of terms, from term 89 on the left to term 12 on the right, are associated with expression and purification, and confirm the validity of the label for research front 17, labeled 'toxin expression and purification.'
- a series of terms at the extreme right, from term 73 on the left to term 62 on the right, are terms associated with warfare, terrorism, and public disasters. These terms are used in papers in the bioterrorism research fronts and help to confirm some of the labels in this group of research fronts.

15.10 Discussion of postal bioterror attacks

The postal bioterror attacks in Fall, 2001, caused a great shock in the specialty of anthrax research. The previous study by Morris, et al, conducted on papers gathered on December 23, 2001, about two months after the attack, showed that 6 papers had already been published in reaction to those attacks. Figure 73 shows a timeline from that study that shows a research front on anthrax bioterrorism. In this diagram the citation links between papers are shown on the timeline to papers cited heavily from the research front. Six papers that appeared after the bioterror attacks are shown and it is noted that these six papers mostly cited a paper by Dixon that covered the treatment of anthrax. Previous to this, papers in the bioterror research front rarely cited Dixon. This indicated that a research front on medical treatment of anthrax was about to emerge.

Looking at Figure 73, which is based on papers gathered on February 25, 2003, about 14 months after the bioterror attacks, the response of the specialty to the postal attacks is evident. There are three additional bioterror related research fronts in the specialty: 1) research front 1 dealing with bioterror incidence management, 2) research front 2 dealing with medical treatment of anthrax, and 3) research front 14 dealing specifically with the postal attacks themselves. Because of the great interest in anthrax research

that was generated after the attacks, a series of anthrax toxin research review papers was generated, which became research front 25 at the top of the timeline.

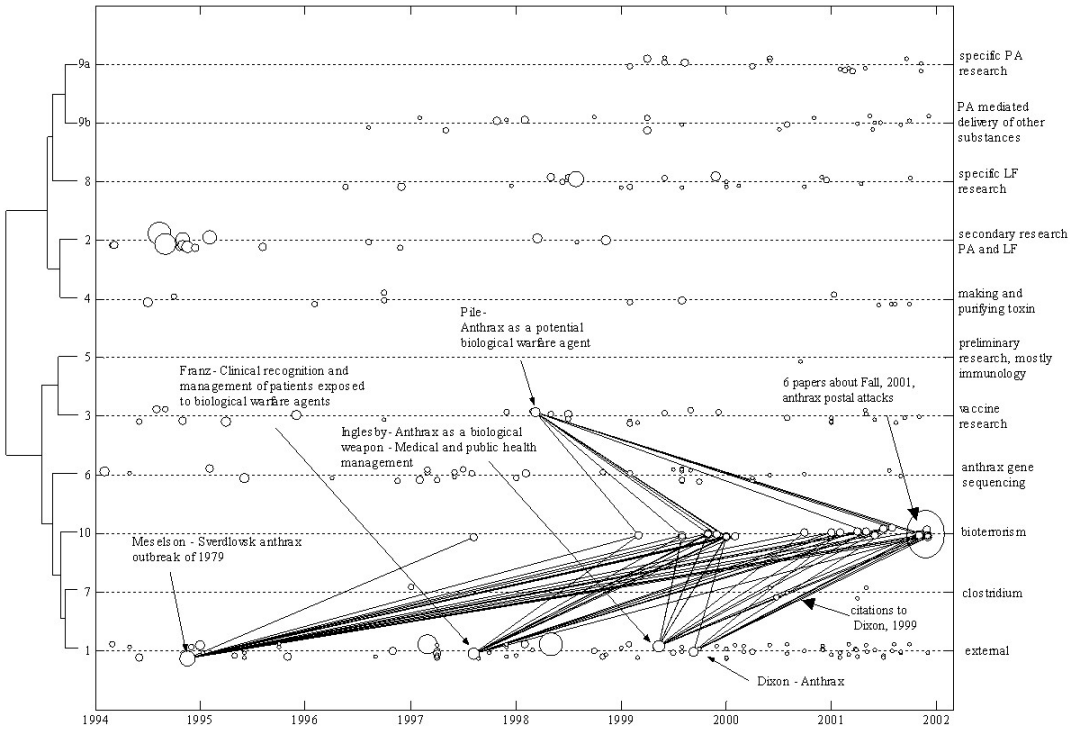


Figure 73. A timeline from an earlier study showing the early effects of the postal bioterror attacks on the literature.

As seen in the timeline of Figure 73, the prediction of a new research front was accurate. There was, however, no anticipation that research fronts on incidence management, the postal attacks themselves, and anthrax toxin review would emerge. Looking at Figure 67 it can be seen that the paper by Dixon on medical treatment of anthrax is heavily cited from the new research front on medical treatment.

This example shows the power of using visualization of data in collections of papers for early detection of emerging research topics. While the new research front on medical treatment was predicted, there was no prediction of the emergence of other important research fronts on the postal attacks, incidence management and anthrax review. The development of these other research fronts probably could have been detected by a program to monitor and analyze the literature periodically after the bioterror attacks.

15.11 Discussion

This chapter has described a case study on anthrax research where the analysis and visualization techniques described in this report have been applied. Given a collection of papers on the topic of anthrax research, the following information was extracted:

- **Research fronts**. A list of research fronts in the collection that comprise subspecialties of research within the anthrax research specialty. The time relation of these research fronts was exposed through the timeline of Figure 66 and from this the active and inactive research fronts were identified and cataloged. The ability to extract this information is important because it allows subject matter experts to quickly and easily explore a specialty and identify currently important subtopics and the literature associated with those subtopics.
- **Reference groups**. The key references within the specialty were identified, clustered into related groups, and their overlapping relation to the specialty's research fronts were shown. The use of these references over time was visualized and the obsolete and current key reference groups were identified. This information was extracted from the research front to reference crossmap, Figure 67, and the reference usage timeline, Figure 68. This analysis is important because it allows subject matter experts to quickly identify the seminal references associated with important subtopics in a specialty. Identification of these references allows the subject matter experts to monitor for new papers that cite these key references, and further allows them to educate themselves on the key elements of these subspecialties by reading the papers corresponding to such references. Furthermore, the overlap of these reference groups with multiple research fronts allows subject matter experts to map how research fronts are related.
- **Reference author groups**. As shown by Figure 69 and Figure 70, the key groups of reference authors in the specialty were identified, clustered into groups and their overlapping relations to research fronts in the specialty were shown. Similar to the information extracted about references, the temporal use of reference authors by research fronts was visualized, showing how groups of references authors, corresponding to "schools of thought" in the specialty, emerged, were used, and became obsolete. Furthermore, this visualization and analysis shows which reference authors are currently being used. The visualization shows experts in each research front as those authors that are well cited by the research fronts.
- **Paper author groups**. As shown in Figure 71, the visualization of paper author usage and clustering allows the identification of teams of researchers, helps to identify prolific authors and also shows which researchers are currently active in the specialty
- **Term groups**. As shown in Figure 72, visualization of terms and groups of terms is very useful for labeling research fronts.

Not shown in this case study are examples of how subject matter experts could use interactive exploration tools provided with the software toolkit described in Chapter 14 to look for specific information in the collection of papers. Nor is the use of the exploration tools provided in interactive web pages described. Additionally, subject matter experts would use written cluster reports generated by the toolkit to explore for specific information in the collection.

Nevertheless, this case study illustrates that the theory and techniques introduced in this report can be used to extract a large amount of useful information about a research specialty from a collection of papers that cover that specialty. Research subtopics, seminal papers, important experts, active research teams, and the relations among them are information that is produced that can be made available to subject matter experts to assess the state of the specialty and make recommendations to planners and research managers.

16. CONCLUSION

16.1 Summary

Motivation. The original motivation for this research was to satisfy a need to explore the structure and dynamics of collections of papers for the purpose of briefing subject matter experts who participate in technology forecasting panels. At that time the state of the art of techniques for analyzing collections of papers consisted of visualizing the papers on a two-dimensional map using multidimensional scaling (MDS), usually performed based on co-citation or Small's similarity. The conventional mental model of a collection of papers, as presented in both the bibliometrics literature and the complex networks literature, is of a large network of papers whose links are citations from one paper to another. In this mental picture there is no separation of papers from references, that is, references are pictured as papers that receive citations and are not considered as entities in and of themselves. Further, as part of the current mental model, the other entities in the collection of papers, authors, journals and terms, when considered as networks at all, are considered as single entity-type networks linked by co-occurrence, e.g., a collaboration network of authors linked by the co-authorship of papers. Another part of the current paradigm for analyzing collections of papers is the pervasive use of co-citation to find links directly between references, or indirectly between papers. To support this assertion, consider a collection of 3103 papers that was gathered that included all papers from the journal *Scientometrics* and the papers that cite them. This collection generally covers the field of bibliometrics and the highly cited references in this collection represent the exemplars to paradigms used within the specialty of bibliometrics. Within this collection, the 6th most cited reference corresponds to Small's discovery paper on co-citation (Small, 1973), the reference corresponding to Kessler's discovery paper on bibliographic coupling (Kessler, 1963) ranks only 54th in number of citations received. This indicates the predominance of co-citation as a metric for clustering over bibliographic coupling, which could serve as an alternative metric.

An examination of the *Scientometrics* paper collection helps to reveal that the current paradigm for bibliometric analysis consists of the following concepts and accompanying exemplars:

- Journal literature is considered a network of papers citing papers.
- Co-citation is the preferred method of establishing links between papers, authors and journals for clustering and mapping.
- MDS maps are the preferred method of visualizing papers, authors and journals.

Working within this paradigm, however, it is practically quite difficult to extract information that can be used directly for the original goal of briefing subject matter experts. The information that subject matter experts would like to review before participating in a technology forecasting panel can be summarized in a few points:

- identification of the key sub-topics in the specialty
- identification of the seminal papers and key references
- identification of experts in the specialty
- identification of centers of excellence (key academic, commercial and government research organizations)
- identification of outside fields contributing loan knowledge to the specialty
- identification of outside fields borrowing loan knowledge from the specialty
- identification of emerging sub-topics
- identification of declining sub-topics

There are many problems with extraction of the desired information about a specialty in the list above when using the current paradigm of bibliometrics. The principal problem is that the current paradigm treats the different entity-types within a collection as separate networks: 1) citation networks of papers, 2) collaboration networks of paper authors, 3) author co-citation networks of reference authors, and 4) journal citation networks. As shown by the work presented here, all of these networks are intimately connected and analysis of any one of them in isolation will inevitably fall short of providing the same amount of information that could be extracted if they are analyzed as a complex interconnected network.

The main mental model in the current bibliometric paradigm, that journal literature is a network of papers citing other papers, has problems as a model for extracting useful information about a research specialty. This model denies that references are used as concept symbols, a crucial limitation to thinking about the underlying processes of literature growth. At best, the current paper citing papers mental model allows the historical tracing of ideas from paper to paper (Garfield et al., 2003), but this "genealogy of ideas" is information that is not particularly important for briefing subject matter experts.

Fortunately, another element in the current bibliometric paradigm, co-citation clustering, so pervasive in current bibliometric practice, can be used to classify both papers and references to find current papers and key references to sub-topics within a specialty, information that is quite useful for subject matter experts. Co-citation clustering and author co-citation clustering are the two most practical and useful of bibliometric techniques being applied today to extract information from journal literature. However, there is little research effort being expended now to investigate the use of other co-occurrence metrics, such as

bibliographic coupling, or to investigate the symbolic representations, as discussed in Chapter 11.1, of co-occurrence groups formed using different co-occurrence metrics.

Another major problem with the current paradigms of bibliometrics is the over-reliance on MDS mapping as a visualization tool. Experiences with using MDS maps for briefing subject matter experts have been uniformly discouraging. MDS maps give a crude view of the structure of the network being mapped, while subject matter experts typically hold much better mental maps of their specialty. A typical reaction of a subject matter expert to an MDS map is "Tell me something I *don't* know." Not only this, but MDS maps typically contain a great number of artifacts, because two dimensional maps cannot preserve complex distance relations among the entities in their original complex many-dimensioned feature space. Subject matter experts view these obvious (given their knowledge of the specialty) inaccuracies with suspicion and turn away from the map as uninformative, inaccurate, and not credible.

Technical review of this research. The work reported here was presented in a logical sequence to explain the proposed mathematical treatment and outline its extensions and applications:

- An entity-relationship model was introduced that explicitly describes the types of entities in a collection of papers and the direct links between them. The correspondence between bibliometric entities and physical entities was explained.
- A method of computing indirect links, based on modeling entities as cascaded bipartite networks and using generalized matrix arithmetic with link weight functions, was discussed. Dyadic links among entities in the collection was explained. The concepts of like and unlike entities, and primary and relative entity-type were explained. A concise mathematical notation, the dyad identifier, was introduced. The dyad identifier notation greatly facilitates the understanding of indirect links and co-occurrence links in the collection.
- Given the model of the paper collection as a collection of bipartite networks, occurrence networks were introduced to mathematically list the links in individual bipartite networks. The use of matrix multiplication to compute indirect links was discussed, and the use of membership matrices, which express entity group memberships, and equivalence matrices, which show correspondence of bibliometric entities to physical entities, was introduced.
- Co-occurrence matrices were introduced to list co-occurrence links among like entities. It was shown that the computation of co-occurrence matrices is equivalent to computing links in a cascade of bipartite networks and that link weight functions and generalized matrix arithmetic could be used accordingly for computing co-occurrence link weights.
- Using occurrence and co-occurrence matrices as a foundation, it was shown that bibliometric distributions can be concisely expressed in terms of these matrices. Dyadic distributions, expressing probability of occurrence of the number of associations of individual entities with

other entity-types, were introduced and shown to occur in two classes, fixed occurrence dyadic distributions, and cumulating occurrence dyadic distributions. Co-occurrence distributions, expressing the probability of the number of co-occurring entities given a pair of like entities, were introduced. Clustering coefficient distributions, which measure the amount of local clustering in a co-occurrence network of like entities, were also introduced.

- A recursive model of occurrence matrix growth and co-occurrence matrix growth was described mathematically and its potential as a tool for modeling growth of the collection of papers was discussed.
- The construction and use of graph theoretic matrices that describe the collection of papers as a paper to cited paper graph was discussed. The derivation of four coupling metrics from the paper graph was discussed. These coupling metrics are: 1) direct citation, 2) co-citation, 3) bibliographic coupling, and 4) longitudinal coupling.
- The direct computation of similarity values from co-occurrence matrices was discussed. These similarity values are used for clustering and visualization of entities for analysis of the paper collection. Methods for fusing similarity values that were derived from two or more different co-occurrence matrices was explained and Small's similarity, a well-known metric for mapping papers from graph theoretic metrics, was explained and generalized.
- The derivation of feature vectors that characterize entities in the pattern recognition sense was introduced. Two types of feature vectors were discussed, 1) occurrence feature vectors, and 2) co-occurrence feature vectors, which are derived from rows of occurrence matrices and co-occurrence matrices respectively. The idea of a characterizing pattern associated with a feature vectors was introduced and discussed.
- Seriation, matrix shading and clustering were explained in the context of the proposed mathematical treatment. It was shown that these operations can be thought of as performing permutations on an occurrence matrix to obtain approximations of a Robinson matrix.
- The visualization of occurrence matrices as a means to understand static and dynamic structures of links within a collection of papers was discussed. Timelines, usage plots, and crossmaps were the three types of visualizations discussed. Several example applications of these visualization techniques were exhibited, timelines showing birth of research fronts and knowledge borrowing, usage plots showing emergence of exemplar references and schools of thought, and crossmaps showing static structure of overlapping links among groups of entities in the collection.
- It was shown that the proposed treatment could be used to efficiently describe many of the analysis methods presently being used in bibliometric analysis. These methods include co-occurrence clustering techniques such as co-citation analysis, latent variable techniques such as Latent Semantic Analysis, and network pruning techniques such as pathfinder analysis.
- A software toolkit that applies the proposed mathematical treatment presented here was presented. This software performs analysis and visualization of collections of journal papers and

patents for the purpose of extracting information about a specialty by analyzing a collection of journal papers covering that specialty. The software has handled paper collections as large as 15,000 papers and makes extensive use of MATLAB's matrix functions and sparse matrix routines.

- A case study was presented, dealing with a collection of 2274 papers covering 60 years of anthrax research. Through analysis and visualization enabled by the proposed mathematical treatment, the growth and specialization of anthrax research through key discoveries was mapped and understood. The birth and obsolescence of sub-topics was mapped, papers were classified into groups by topic (research fronts.) Key references, experts, and active research teams were identified. Additionally, the overlapping correspondence of groups of related references, reference authors, and paper authors to the specialty's research topics was visualized and the emergence and obsolescence of these groups over time was visualized. This case study clearly demonstrates the usefulness of the mathematical treatment proposed here.

As listed above, the mathematical treatment is general, easily usable, and can potentially be exploited by a great number of analytical, statistical and visual techniques for extracting required information and knowledge from a collection of papers.

16.2 Significance of this research

It is important to review the significance of the research presented here, particularly in relation to current techniques for both bibliometric analysis and analysis of complex networks:

Modeling of multiple entity-type networks. In the context of both complex network analysis and bibliometric analysis, there are presently no useful techniques for modeling and simulating networks that contain more than one entity-type. While there is some presentation of analysis techniques for bipartite networks (Dorogovtsev & Mendes, 2002), this analysis is not pursued in a general sense. Recent work by Borner, Maru and Goldstone (2004) in bibliometrics presents a model of simultaneous evolution of networks of authors and papers. The model does not address the mathematical expression of the links between authors, papers and references, but rather focuses on the evolution of two distinct and separate networks, the author network and the paper network, as the result of some underlying process. *In contrast*, t*he mathematical treatment here allows the simultaneous simulation of all entity-types in a collection of papers or complex network*. As an example Morris (2004) modeled and simulated the simultaneous growth of papers and references in a collection of papers, producing simulations that matched a large number of characteristics of actual paper collections: 1) the paper per reference distribution, 2) the reference per paper distribution, 3) the co-citation distribution, 4) the bibliographic coupling distribution, 5) the bibliographic coupling clustering distribution, and the temporal distribution of citation counts. This model, using a

cumulative advantage process (Price, 1976; Simon, 1955) applied to the growth of a paper reference matrix as modeled in Chapter 7.2, was able to show the key role of highly cited exemplar references in the production of the dense network of links in co-citation networks of references and bibliographic coupling networks of papers. That work, combined with the mathematical treatment proposed here, can easily be generalized to include general simulation of the all entity-types in the collection of papers simultaneously.

Quantifying indirect links. There is very little work presented in the complex networks literature or bibliometrics literature on computing the weights of links between entities. Indeed, in most complex networks research, links are simply considered as unweighted. Most research on complex networks is done statistically, principally modeling the distribution of the number of links per entity (Albert & Barabasi, 2002). There is very little work in the information retrieval and bibliometrics literature on quantifying indirect links, where most links are derived from simple co-occurrence counts. In fact, the concept of indirect links is probably not known in either of these fields. *The mathematical treatment proposed here provides a systematic, general, and easy to implement method of computing indirect links based on link weight functions and matrix arithmetic.* The calculation method is general enough to incorporate most or all of the candidate methods of link computation for indirect links: matrix multiplication, the overlap function, and the inverse Minkowski metric. The method is easily implemented in software and is completely amenable to sparse matrix techniques developed for matrix multiplication.

Characterizing growth of entities and links. The matrix formulation of the growth of occurrence and co-occurrence matrices is a possible basis for describing the general growth of complex networks that can be described by the entity-relationship model outlined in Chapter 2. This model, along with the use of cascaded bipartite networks as described in Chapter 3, naturally leads to the general mathematical formulation of network growth using recursive matrix equations as described in Chapter 7. This recursive formulation simply describes the growth of occurrence matrices in the network and further describes the resulting growth of co-occurrence matrices in a useful way that makes obvious the difference between static links and cumulating links in the network. *The recursive matrix equations introduced here allow direct efficient modeling and simulation of the growth of a complex multiple entity-type network.* The recursive formulation can be directly applied to network simulation and studies, as was done by Morris (2004) to model and simulate the manifestation of the growth of a specialty in its literature. The recursive matrix growth model suggests methods for efficient storage and computational computer algorithms when analyzing large networks.

Characterizing classes of distributions. The characterizations of distributions discussed in Chapter 6 provide a unique view of network distributions that has never been investigated. The efficient mathematical notation and nomenclature, the entity-relationship model, and the concepts of static and cumulating links, lead to a general and symmetric view of bibliometric distributions. To date, research on distributions in

bibliometrics and complex networks has focused almost exclusively on power-law distributions (Albert & Barabasi, 2002; Fairthorne, 1969). There has been some study of static occurrence distributions in the literature in the context of collaboration networks (Newman, Watts, & Strogatz, 2002). Nevertheless, *the mathematical treatment introduced here allows the efficient general description of dyadic distributions and allows classification of those distributions into static occurrence and cumulating occurrence distributions. The mathematical treatment further allows the general study of co-occurrence distributions and clustering coefficient distributions.* As shown by Morris (2004), static dyadic distributions, almost totally neglected in the current research on bibliometrics and complex networks, must be modeled correctly in order to study and simulate growth of the networks in a collection of papers.

Standardization of mathematical characterizations. The simple dyad identifier notation introduced in this mathematical treatment greatly facilitates the modeling and analysis of collections of papers. The number of bipartite networks, distributions, and co-occurrence networks in a collection of papers makes the description of the networks exceedingly cumbersome without an efficient notation. The setup of calculations of indirect links using matrix arithmetic is reduced to triviality when dyad identifier notation is used. *The concept of like and unlike entities, primary and relative entity-types, and most of all, the dyad identifier notation, makes description of the collection of papers very efficient and elegantly simple.* This contribution to network analysis will greatly facilitate communication of ideas and should encourage progress in research in complex networks.

Calculation of similarities. The treatment presented here standardizes the calculation of similarities from co-occurrence matrices as shown in Chapter 9. More importantly, *the treatment presented here shows how to calculate similarities from links that are not based on simple co-occurrence counts, but generalized to links based on general link weight functions. Additionally, the treatment facilitates the fusion of similarities from multiple types of links for analysis* as shown in Chapter 9.2 and Chapter 9.3.

Clustering, seriation and matrix shading. It was shown in Chapter 11 that it is possible to show the results of clustering as permutation of occurrence matrices to approximate a Robinson matrix that exposes structures in the links in a bipartite network. It was shown that matrix shading, clustering and seriation are related in their effects on rearranging an occurrence matrix. *The mathematic treatment introduced here consolidates understanding of matrix shading, seriation, and clustering of entities in the paper collection and facilitates the visualization of links among groups of entities in the collection*. This allows understanding of the relations among groups of like entities, but further allows analysis of relations among groups of unlike entities, an innovation in analysis not previously used to analyze collections of papers.

Introduction of useful visualization techniques. To date, the prevailing paradigm of visualization of collections of papers has been the two dimensional MDS map. Software to generate these maps is readily available in statistics programs such as SAS. As discussed above, MDS maps are capable of visualizing crude structure, but often produce artifacts that mislead the analysts using them to wrong conclusions. They are also not capable of effectively communicating overlapping relations among groups of entities, and dynamic changes in the structure of links in the collection of papers. The visualization methods introduced in Chapter 6, based on displaying the structure of occurrence matrices, are not focused on mapping entities, as MDS maps do, but are focused on mapping links between groups of entities. *The visualization techniques introduced here produce efficient displays of complex and overlapping links between groups of entities, and also produce useful displays of trends and emerging events in the structure of links in the collection of papers.* A wide variety of useful displays can be produced and related to the entity-relationship model of the collection of papers. These displays can be easily related to each other and allow an analyst to see the complex network of links in the collection of papers from many interlocking perspectives.

Incorporation of existing analysis into the proposed mathematic treatment. As shown in Chapter 13, the proposed mathematical treatment can be used to express a great number of existing analysis techniques presently being used to analyze collections of papers. Among these techniques are 1) co-occurrence clustering techniques such as co-citation analysis and author co-citation analysis, 2) latent variable and modal analysis techniques such as latent semantic analysis, factor analysis, and hub and authorities analysis, 3) feature vectors to characterize individual entities, and 4) network pruning techniques such as pathfinder analysis. *The mathematical treatment proposed here allows existing bibliometric analysis techniques to be incorporated into the same mathematical framework for comparison and generalization*. For example, it is easy to show, using the mathematical treatment introduced here, that latent semantic analysis can be easily applied to links between papers and references, or to the links between paper authors and reference authors. This generalization of analysis techniques, facilitated by the mathematical treatment, provides many new, easily explored ideas for further research in the field.

Simulation of cascaded bipartite networks. *Using Yule models the mathematical treatment presented here allows a very general method of simulating growing cascaded bipartite networks such as collections of journal papers.* Morris (2004) showed that a modified Yule model can accurately describe the growth of paper to reference networks in a specialty, while Goldstein, Morris, and Yen (in print) show that another type of modified Yule model can be used to simulate the manifestation of research teams in a specialty. Yule models, though not discussed in this report, are well suited to growth models of cascaded bipartite networks because they model linear growth in the number of entities and preferential connection, two key characteristics of such networks.

16.3 Future Research

The work here can be extended with the following research and applications:

- **Extension to other applications**. There are several data sources, similar to collections of papers that could be adapted to entity-relationship models and the mathematical treatment presented here. Specifically, three examples are:

    - **Press clippings**. Press clippings are topic-specific reports culled from general press reports that are analogous to reports provided by a professional clipping service. An example of a press clipping collection would be press reports on terrorist events. Press reports must undergo an entity extraction processing step to build the report database in entity-relationship form.
    - **Patents.** These are issued by governments, and are documents describing inventions and granting exclusive rights to exploit those inventions. Patent abstract data is readily available from a number of sources and is used to monitor technology for competitive intelligence, technology forecasting, and other business purposes.
    - **Film databases**. Film databases, covering thousands of films, are available free from the Internet. These records have been extensively studied in the complex networks literature and contain information on each film such as title, actors, director, producer, and release date. The data is readily converted to an entity-relationship model for study using the mathematical treatment presented here.
- **Investigation of matrix growth equations**. The matrix growth equations introduced in Chapter 7 are potentially useful to model modes of growth in a complex network and develop methods of monitoring for both trends and discontinuous events. Further work to develop applications for the recursive formulation should be pursued.
- **Development of clustering algorithms** Clustering algorithms are the weakest link in analysis of collections of papers and other complex networks. This is especially true for co-occurrence networks based on cumulating links, where the distribution of co-occurrences is often highly skewed by the power-law distributions of occurrences with entities of the relative entity-type. There is a need to investigate methods of fusing information from many sources to obtain robust clusters of entities.
- **Interactive visualization techniques**. Initial work has been conducted to provide an interactive interface to visualizations of Chapter 6 using Web based tools programming tools such as Javascript and ASP. For example, at samorris.ceat@okstate.edu/web/case_studies a number of case studies can be viewed that allow a subject matter expert to use timelines, usage maps, and crossmaps to access papers, references, paper authors, reference authors, and paper journals, and reference journals associated with specific research fronts. Furthermore, animations of growth of

research fronts can be executed by the user. These interactive visualizations should be explored to find the best method of allowing subject matter experts to explore a specialty through its literature.

- **Generalization to include complex network theory**. The entity-relationship model and the mathematical treatment here need to be generalized to work within the framework of current complex network theory research. This promises to extend the current research in complex network theory to networks with an arbitrary number of entity-types, in contrast with the current complex network models that have only one or occasionally two entity-types.

16.4 Concluding remarks

The mathematical treatment introduced here provides an easily understood and easily implemented way of working with collections of papers as a complex network. The ability it gives of computing links between any arbitrarily selected entity-types greatly simplifies the analysis of collections of papers. Much of existing theory on collections of papers: 1) similarity computations, 2) bibliometric distributions, 3) feature vectors, 4) clustering and seriation, and many other analysis techniques, readily fit into the framework of the mathematical treatment presented here. Not only this, but once these techniques are adapted to the framework of the proposed mathematical treatment, there are many extensions to those techniques that become apparent and that warrant investigation as possibly useful. Viewed this way, it appears the proposed mathematical technique may be able to function as a framework within which research on analysis of collections of papers and other multiple entity-type complex networks can be investigated.

BIBLIOGRAPHY

Ahlgren, P., Jarneving, B., & Rousseau, R. (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. *Journal of the American Society for Information Science and Technology, 54*(6), 550-560.

Albert, R., & Barabasi, A. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47-97.

Asnake, B. (2003). *Automatic scientific literature classification using multiple information sources for data mining purposes.* Unpublished Master of Science thesis, Oklahoma State University, Stillwater, Oklahoma, USA.

Bar-Joseph, Z., Gifford, D. K., & Jaakola, T. S. (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics, 17*(S1), S22-S29.

Beaver, D. D. (1978). Studies in scientific collaboration. Part 1. The professional origins of scientific co-authorship. *Scientometrics, 1*, 65-84.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review, 37*(4), 573-595.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press.

Bhatnagar, R., & Batra, S. (2001). Anthrax toxin. *Critical Reviews in Microbiology, 27*(3), 167-200.

Bookstein, A. (1990). Informetric distributions, part I: unified overview. *Journal of the American Society for Information Science and Technology, 41*(5), 368-375.

Borner, K., Chen, C., & Boyack, K. W. (2002). Visualizing knowledge domains. *Annual Review of Information Science and Technology, 37*, 179-255.

Borner, K., Maru, J. T., & Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Science of the United States, 101*(suppl. 1), 5266-5273.

Boyack, K. W., Wylie, B. N., & Davidson, G. S. (2002). Domain visualization using VxIsight for science and technology management. *Journal of the American Society for Information Science and Technology, 53*(9), 764-774.

Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science and Technology, 42*(4), 233-251.

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

Brower, J. C., & Kile, K. M. (1988). Seriation of an original data matrix as applied to paleoecology. *Lethaia, 21*, 79-93.

Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics, 52*(1), 3-12.

Callon, M., Courtial, J. P., & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research - the case of polymer chemistry. *Scientometrics, 22*(1), 155-205.

Chen, C. (1998). Bridging the gap: the use of pathfinder networks in visual navigation. *Journal of Visual Languages and Computing, 9*, 267-286.

Chen, C., Cribbin, T., Macredie, R., & Morar, S. (2002). Visualizing and tracking the growth of competing paradigms: two case studies. *Journal of the American Society for Information Science and Technology, 53*(8), 678-689.

Chen, C. M., & Morris, S. A. (2003, October 19-21, 2003). *Visualizing evolving networks: Minimum spanning trees versus Pathfinder networks.* Paper presented at the IEEE Symposium on Information Visualization, Seattle, Washington.

Chen, P. (1976). The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems, 1*(1), 9-36.

Cios, K. J., Pedrycz, W., & Swiniarski, R. (1998). *Data mining methods for knowledge discovery*. Boston: Kluwer Academic.

Crane, D. (1980). An exploratory study of Kuhnian paradigms in theoretical high energy physics. *Social Studies of Science, 10*, 23-54.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391-407.

Ding, Y., Chowdhury, G. G., & Foo, S. (2000). Journal as markers of intellectual space: journal co-citation analysis of information retrieval area, 1987-1997. *Scientometrics, 47*(1), 55-73.

Doreian, P. (1988). Testing structural-equivalence hypotheses in a network of geographic journals. *Journal of the American Society for Information Science, 39*(2), 79-85.

Dorogovtsev, S. N., & Mendes, J. F. F. (2002). Evolution of networks. *Advances in Physics, 51*(4), 1079-1187.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification* (2nd ed.). New York: Wiley.

Fairthorne, R. A. (1969). Empirical hyperbolic distributions (Bradford-Zipf-Mandelbrot) for bibliometric description and prediction. *Journal of Documentation, 25*(4), 319-343.

Garfield, E. (1994). Research fronts. *Current Contents, 41*, 3-7.

Garfield, E., Pudovkin, A. I., & Istomin, V. S. (2003). Mapping the output of topical searches in the Web of Knowledge and the case of Watson-Crick. *Information Technology and Libraries, 22*(4), 183-187.

Goldstein, M. L., Morris, S. A., & Yen, G. (2004). Problems with fitting to the power-law distribution. *European Physical Journal B, 41*, 255-258.

Goldstein, M. L., Morris, S. A., & Yen, G. G. (in print). A group-based model for bipartite author-paper networks. *Physical Review E, cond-mat/0409205*.

Gordon, A. D. (1999). *Classification* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Hargens, L. L. (2000). Using the literature: reference networks, reference contexts, and the social structure of scholarship. *American Sociological Review, 65*(6), 846-865.

Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: a geometrical analysis of similarity measures. *Journal of the American Society for Information Science and Technology, 38*(6), 420-442.

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

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

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills: Sage Publications.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2d ed.). Chicago: University of Chicago Press.

Lambert, D. (1992). Zero inflated Poisson regression, with application to defects in manufacturing. *Technometrics, 34*(1), 1-14.

Lenstra, J. K. (1974). Clustering a data array and the traveling salesman problem. *Operations Research, 22*, 413-414.

Leydesdorff, L. A. (1995). *The challenge of scientometrics: the development, measurement, and self-organization of scientific communications.* Leiden: DSWO Press, Leiden University.

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

McCain, K. W. (1990). Mapping authors in intellectual space: a technical overview. *Journal of the American Society for Information Science and Technology, 41*(6), 433-443.

McCain, K. W. (1991). Mapping economics through the journal literature: an experiment in journal cocitation analysis. *Journal of the American Society for Information Science, 42*(4), 290-296.

McCain, K. W. (1998). Neural networks research in context: a longitudinal journal cocitation analysis of an emerging interdisciplinary field. *Scientometrics, 41*(3), 389-410.

Morris, S. A. (2004). Manifestation of emerging specialties in journal literature: a growth model of papers, references, exemplars, bibliographic coupling, co-citation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology, in press*.

Morris, S. A., Asnake, B., & Yen, G. (2003). Optimal dendrogram seriation using simulated annealing. *Information Visualization, 2*(2), 95-104.

Morris, S. A., DeYong, C., Wu, Z., Salman, S., & Yemenu, D. (2002). DIVA: a visualization system for exploring document databases for technology forecasting. *Computers and Industrial Engineering, 43*(4), 841-862.

Morris, S. A., Wu, Z., & Yen, G. (2001, July 14-19). *A SOM mapping technique for visualizing documents in a database.* Paper presented at the International Joint Conference on Neural Networks Proceedings, Washington D. C.

Morris, S. A., & Yen, G. (2004). Crossmaps: visualization of overlapping relationships in collections of journal papers. *Proceedings of the National Academy of Science of the United States, 101*(suppl. 1), 5291-5296.

Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Timeline visualization of research fronts. *Journal of the American Society for Information Science and Technology, 54*(5), 413-422.

Naranan, S. (1971). Power law relations in science bibliography- a self-consistent interpretation. *Journal of Documentation, 27*(2), 83-97.

Newman, M. E. J., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America, 99*(suppl 1), 2566-2572.

Packer, C. V. (1989). Applying row-column permutation to matrix representations of large citation networks. *Information Processing & Management, 25*(3), 307-314.

Persson, O. (1994). The intellectual base and research fronts of JASIS 1986-1990. *Journal of the American Society for Information Science and Technology, 45*(1), 31-38.

Price, D. (1965). Networks of scientific papers. *Science, 149*(3683), 510-515.

Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science, 27*(5-6), 292-306.

Redner, S. (1998). How popular is your paper? an empirical study of the citation distribution. *European Physical Journal B, 4*(2), 131-134.

Robinson, W. A. (1951). A method for chronologically ordering archaeological deposits. *American Antiquity, 16*(4), 1350-1362.

Salton, G. (1971). *The SMART retrieval system; experiments in automatic document processing*. Englewood Cliffs: Prentice-Hall.

Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer.* Reading: Addison-Wesley.

Schagrin, M. L. (1963). Resistance to Ohm's Law. *American Journal of Physics, 31*, 536-547.

Schvaneveldt, R. W., Dearholt, D. W., & Durso, F. T. (1988). Graph theoretic foundations of pathfinder networks. *Comput. Math. Applic., 15*(4), 337-345.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. *The Psychology of Learning and Motivation, 24*, 249-284.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science, 43*(9), 628-638.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika, 42*, 425-440.

Small, H. (1973). Cocitation in scientific literature - new measure of relationship between 2 documents. *Journal of the American Society for Information Science, 24*(4), 265-269.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*, 327-340.

Small, H. (1997). Update on science mapping: creating large document spaces. *Scientometrics, 38*(2), 275-293.

Turnbull, P. C. B. (1991). Anthrax vaccines: past present and future. *Vaccine, 9*, 533-539.

White, H. D. (2001). Authors as citers over time. *Journal of the American Society for Information Science and Technology, 52*(2), 87-108.

White, H. D. (2003a). Author cocitation analysis and Pearson's r. *Journal of the American Society for Information Science and Technology, 54*(13), 1250-1259.

White, H. D. (2003b). Pathfinder networks and author cocitation analysis: a remapping of paradigmatic information scientists. *Journal of the American Society for Information Science and Technology, 54*(5), 423-434.

White, H. D., & Griffith, B. C. (1981). Author cocitation: a literature measure of intellectual structure. *Journal of the American Society for Information Science, 32*(3), 163-172.

White, H. D., & McCain, K. W. (1989). Bibliometrics. *Annual Review of Information Science and Technology, 24*, 119-186.

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

Zhu, D., & Porter, A. (2002). Automated extraction and visualization of information for technological intelligence and forecasting. *Technological Forecasting and Social Change, 69*, 495-506.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Reading: Addison-Wesley.

VITA

Steven Allen Morris

Candidate for the Degree of

DOCTOR OF PHILOSOPHY

Thesis: UNIFIED MATHMATICAL TREATMENT OF COMPLEX CASCADED BIPARTITE NETWORKS: THE CASE OF COLLECTIONS OF JOURNAL PAPERS

Major Field: Electrical Engineering

Biographical:

Education: Received a Bachelor of Science degree in Electrical Engineering from Tulsa University, Tulsa, Oklahoma, in May, 1983. Received a Master of Science degree in Electrical Engineering from Tulsa University, Tulsa, Oklahoma in May, 1987. Completed the requirements for the Doctor of Philosophy with a major in Electrical Engineering at Oklahoma State University in May of 2005.

Experience: Research Scientist, Amoco Production Company, Tulsa, Oklahoma (1985-1990). Senior Research Scientist, Amoco Production Company, Tulsa Oklahoma (1990-1999). Research Associate, Oklahoma State University, Stillwater, Oklahoma (2001-2005). Adjunct Professor, Oklahoma State University, Stillwater, Oklahoma (2001-2004).

Professional Memberships:
The Institute of Electrical and Electronic Engineers.

Name: Steven Morris Date of Degree: May, 2005

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: UNIFIED MATHEMATICAL TREATMENT OF COMPLEX CASCADED BIPARTITE NETWORKS: THE CASE OF COLLECTIONS OF JOURNAL PAPERS

Pages in Study: 168 Candidate for the Degree of Doctor of Philosophy

Major Field: Electrical Engineering

A mathematical treatment is proposed for analysis of entities and relations among entities in complex networks consisting of cascaded bipartite networks. This treatment is applied to the case of collections of journal papers, in which entities are papers, references, paper authors, reference authors, paper journals, reference journals, institutions, terms, and term definitions. An entity-relationship model is introduced that explicitly shows direct links between entity-types and possible useful indirect relations. From this a matrix formulation and generalized matrix arithmetic are introduced that allow easy expression of relations between entities and calculation of weights of indirect links and co-occurrence links. Occurrence matrices, equivalence matrices, membership matrices and co-occurrence matrices are described. A dynamic model of growth describes recursive relations in occurrence and co-occurrence matrices as papers are added to the paper collection. Graph theoretic matrices are introduced to allow information flow studies of networks of papers linked by their citations. Similarity calculations and similarity fusion are explained. Derivation of feature vectors for pattern recognition techniques is presented. The relation of the proposed mathematical treatment to seriation, clustering, multidimensional scaling, and visualization techniques is discussed. It is shown that most existing bibliometric analysis techniques for dealing with collections of journal papers are easily expressed in terms of the proposed mathematical treatment: co-citation analysis, bibliographic coupling analysis, author co-citation analysis, journal co-citation analysis, Braam-Moed-vanRaan (BMV) co-citation/co-word analysis, latent semantic analysis, hubs and authorities, and multidimensional scaling. This report discusses an extensive software toolkit that was developed for this research for analyzing and visualizing entities and links in a collection of journal papers. Additionally, an extensive case study is presented, analyzing and visualizing 60 years of anthrax research.. When dealing with complex networks that consist of cascaded bipartite networks, the treatment presented here provides a general mathematical framework for all aspects of analysis of static network structure and network dynamic growth. As such, it provides a basic paradigm for thinking about and modeling such networks: computing direct and indirect links, expressing and analyzing statistical distributions of network characteristics, describing network growth, deriving feature vectors, clustering, and visualizing network structure and growth.

ADVISOR'S APPROVAL_____Dr. Gary Yen_____