III. CITATION ANALYSIS

Citations are frozen footprints on the landscape of scholarly achievement.

Cronin

III.0. INTRODUCTION

A scientific paper does not stand alone : it is embedded in the literature of the subject. The nature of this embedding is specified by the use of footnotes and/or reference lists. The fact that a document is mentioned in a reference list indicates that in the author's mind there is a relationship between a part or the whole of the cited document and a part or the whole of the citing document. Citation analysis is that area of informetrics which deals with the study of these relationships. The main tool for this is a citation index : this is an ordered list of cited documents, each accompagnied by a list of citing documents. Indeed, citation analysis might conceivably not have emerged as a serious academic issue had not the commercial development of citation indexing proved so successful.

According to Zunde (1971) there are three main application areas in citation analysis :

1) qualitative and quantitative evaluation of scientists, publications and scientific institutions;
2) modelling of the historical development of science and technology;
3) information search and retrieval.

A general reference for this part is Garfield's book (1979a). A recent review on citation practices can also be found in Todres (1986), showing the great interest for citation analysis that exists in the Soviet Union.

III.1. CITATION INDEXING

III.1.1. References and citations

Scientific tradition requires, at least since the 19th century, that scientists writing articles refer to earlier articles which relate to the theme of the paper. These references are supposed to identify those earlier researchers whose concepts, methods, equipment, etc. inspired or were used by the author in developing his or her own article. (For a more detailed analysis of citers' motivations, see Chapter III.2).

If one wishes to be precise, one should distinguish between the notions '*reference*' and '*citation*'. If paper R contains a bibliographic note using and describing paper C, then R contains a reference to C and C has a citation from R (Price (1970)). Stated otherwise, a reference is the acknowledgement that one document gives to another, while a citation is the acknowledgement that one document receives from another. So, 'reference' is a backward-looking concept, while 'citation' is a forward-looking one (see Fig.III.1.1).

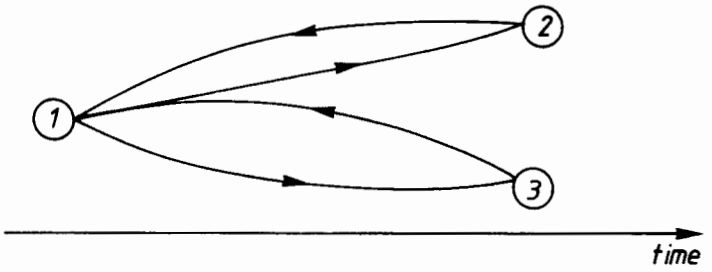

Fig.III.1.1

⟵ : means 'refers to' or 'cites'

⟶ : means 'is cited by'

(1) is a reference from the reference lists of (2) and (3);
(2) and (3) contain citations of (1).

Although most authors are not so precise in their usage of both terms, we agree with Price (1970) that using the words 'citation' and 'reference' interchangeably is a deplorable waste of a good technical term.

III.1.2. The principle of citation indexing

The practice of giving references has been used to construct so-called citation indexes. More precisely, *citation indexing* is based on the idea that an author's references to previously recorded information identify much of the earlier work that is pertinent to the subject of this present document. A citation index is then a structured list of all the references in a given collection of documents.

The first practical application of this principle did not occur in library and information sciences but in legal practice. Indeed, Shepard's Citations, a legal reference tool, has been in use in the United States since 1873. To use Shepard's Citations, a lawyer must first locate a previous decision relating to his current case. He does this by consulting a digest which will provide him with the case number for any given decision. The lawyer then looks up the case number in Shepard's Citations and finds all the subsequent citing cases. From this information, he or she can determine whether the original decision was affirmed.

To all intents and purposes the *Science Citation Index (SCI)*, the *Social Science Citation Index (SSCI)* and the *Arts & Humanities Citation Index (A&HCI)* are the only citation indexes in practical use for scientific purposes. All three are distributed on a commercial basis by the Institute for Scientific Information (ISI), Philadelphia, USA, and are edited under the direction of the man whose idea it was to launch this large enterprise, namely Eugene Garfield.

To set the record straight, Garfield was neither the first to think of the idea of citation indexing for literature retrieval nor the first to apply it. The Institute of Electrical Engineers, for instance, had a citation index in the beginning of the century, which was abandoned in 1922 (Garfield (1971)).

In addition to the citation indexes compiled by ISI and Garfield, there have been other efforts at making citation indexes. However, most of these have been experimental in nature, extremely narrow in their coverage, or published on a one-time basis (Weinstock (1971)).

The first Science Citation Index (SCI), which was published in 1963, covered the literature of the calender year 1961. It covered 613 journals, contained 1.4 million citations and required five volumes. From this first SCI those items dealing with genetics were selected and likewise published separately as the Genetics Citation Index. In 1986 ISI published a retrospective SCI covering the period 1955-1964; the edition covering 1945-1954 appeared in 1988.

III.1.3. Description and use of the Science Citation Index and the Social Science Citation Index

The Science Citation Index and the Social Science Citation Index provide indexes to the contents of every issue published during a calender year of approximately 5600 selected journals. As the composition and possible uses of both indexes are very similar, we will only describe the SCI. The journals which are covered are called 'source journals' and the items they contain are called 'source items'. All original articles and most other useful items such as editorials and letters are processed (see Table III.1.1). Although previously included, book reviews - except those appearing in Nature and Science - have not been used as source items since 1969.

Table III.1.1. Definitions of journal items processed by the Institute for Scientific Information (ISI)

Chronologies - articles that mainly contain lists of events in the sequence in which they occurred.
Corrections, additions - corrections of errors found in articles or additions of information to articles that were previously published and that have been made known after these articles were published.
Discussions, conferences - items in which one or more persons pass comment on a paper, case or topic.
Editorials, interviews - articles that give opinions of persons, groups or organisations.
Individual items - articles focusing on the life of a person and articles that are tributes to or commemorations of a person, for example, obituaries and short biographies.
Letters - contributions or correspondence from the readers to the journal editor concerning previously published material.
Meeting abstracts - general summations of completed papers that were or will be presented at a symposium or conference.
Notes, brief reports, communications - technical comments shorter than an article and restricted in scope.
Proceedings papers - complete papers that were or will be presented at a symposium or conference.
Research reports, papers - articles reporting the results of original work. Most primary research articles fall into this category.
Reviews, bibliographies - critical or analytical examinations of material previously published. Review articles may draw profound conclusions but usually do not include new research data. Bibliographical lists, often with descriptive or critical notes, of writings relating to a particular subject also fall into this category.

All foreign language (i.e. non-English) titles are translated into English and all references, whether as footnotes or provided in a bibliography, are processed.

The SCI originally consisted of three separate but related indexes. These are the *Citation Index*, the *Source Index* and the *Permutern Subject Index*. Since 1976 these three indexes have been augmented with a fourth : the *Journal Citation Reports (JCR)*.

The Citation Index is arranged alphabetically by cited author. An entry for a cited item contains the first author's name and initials (when available), the year the cited item was published, and the name of the publication in which the cited item appeared along with its volume and page number. When there is more than one cited item for an author, these are arranged chronologically.

Table III.1.2. Example of citation data (taken from the SCI, ed. 1988)

EMBRECHTS P						
** IN PRESS SUBEXPONENT						
STAM AJ	J MATH ANAL			130	493	88
78 J LONDON MATH SOC		17	102			
OMEY E	J LONDON MATH			37	105	88
79 Z WAHRSCHEINLICHTKEIT		49	335			
PAKES AG	MATH PROC C			103	351	88
WILLEKEN E	J APPL PROB			25	413	88
' ' '	PROB TH REL			77	567	88
80 J AUSTR MATH SOC A		29	243			
. . .						

The source items citing a particular reference are arranged alphabetically by source author immediately under each reference line. The source item line contains the citing author's name, the name of the publication in which the citing item appeared and volume, page and the publication year. There is also a code indicating whether the citing item was an article, abstract, editorial, etc. References (cited items) may be from any year in recorded history; citing items, however, are always from the current year. In the Citation Index only the first author is shown for the cited and citing items.

A separate section of the Citation Index is used for anonymous items. These items are arranged alphabetically by the titles of the cited publications. Another separate section within the Citation Index contains a Patent Citation Index. This is a listing of all patents which have been cited

in any journal covered by the SCI. The Patent Citation Index is arranged in numerical order by patent number and usually provides, in addition to the patent number cited, the year of issuance, inventor and country.

The Source Index is arranged alphabetically by source item author. Entries provide all co-authors, the full title of the source item, journal title, volume, issue, page, year, type of item and number of references in the bibliography of the source item. Also provided is an accession number : this is the code by which the source journal is filed at ISI.

Within the Source Index there is a separate section called the 'Corporate Index'. In the Corporate Index, all of the source items processed are listed alphabetically by author under the name of the organisation where the work was performed. If more than one organisation was involved, an entry is created for each organisation.

The third major index contained in the SCI is the so-called Permuterm Subject Index (PSI). Here 'Permuterm' is a contraction of the phrase 'permuted terms'. To produce the PSI, a computer is used to form all ordered pairs of significant words within each title of every item included in the Source Index. Thus, for a title containing n significant words, there will be n(n-1) pairs. With this system, every significant word takes a turn at being the primary term as well as being a co-term.

A general reference for this section is Weinstock (1971) and the introductory pages of the SCI. We postpone a description of the Journal Citations Reports until Chapter III.5.

III.1.4. The A&HCI and the online versions of the SCI, the SSCI and the A&HCI

Unlike the SCI and the SSCI, the Arts & Humanities Citation Index indexes 'implicit' citations too. These occur when an article refers to and substantially discusses a specific work but does not formally cite it. Reproductions of works of art and music scores are also considered as implicit citations, with a code indicating that the cited work is an illustration.

In the Arts and Humanities Index there is also the problem of whimsical and other types of inadequate titles. To deal with this problem, the Permuterm Subject Index of the A&HCI indexes an article as if its title contained the name, place or thing it is about.

The disciplines covered by the A&HCI include literature, languages, history, philosophy, religion, classical studies, fine arts and architecture, music and the performing arts of dancing, drama, film, radio and television.

For most scientific investigations the use of the online versions of these citation indexes is more appropriate than that of the printed versions.

The online version of the SCI is called 'SciSearch'. 'Social SciSearch' is the online version of the SSCI and 'Arts & Humanities Search' is the online version of the A&HCI. From 1988 on the Science Citation Index has also been available on CD ROM : discs are issued quarterly, with a complete accumulation available at the end of each year. The CD ROM version of the Social Science Index appeared for the first time in 1989. The compact disc edition of the SCI has been reviewed by Tseng et al. (1988). Although a part of the SCI and the SSCI, the Journal Citation Reports are not available online or on disc.

III.1.5. Deficiences of subject indexes versus citation indexes

Articles cite earlier items which describe related work, concepts, methods, etc. The subject of the cited articles relates to the subject of the citing articles, so that the citation linkages indicate subject relationships. As such, citation indexes are powerful tools for retrieval because the authors of the papers which a searcher knows to be relevant usually make a relevance judgement on the paper they cite that is more useful to the searcher than comparable relevance judgements made by the average indexer (see, however, Chapter III.2 about various reasons why authors cite other authors). Another reason why citation indexes are so useful is their independence of topic-descriptors, avoiding the imprecision and inconsistency inherent in the use of such topic-descriptors (Garfield (1974)).

According to Weinstock (1971), citation indexes have the following advantages over traditional subject indexes :

1. Traditional indexes are no longer able to deal comprehensively with the growing volume of scientific literature on a timely basis.

The SCI's method of obtaining comprehensive coverage of the literature is based on Bradford's law (see Part IV on informetric models). This law states, intuitively, that a small percentage of journals account for a calculably large percentage of the significant journals in any given field of science. Moreover, this concentration of information in relatively few journals is characteristic, not only for the individual discipline, but also for scientific literature as a whole (Price (1963)). So, if ISI choses its journals in a proper way, they are sure to cover the most important journal literature from all over the world. For an account of how ISI actually selects journals for coverage, the reader is referred to Garfield (1985) and Allee (1988), and for a criticism on this selection policy to Scanlan (1988). In particular, the compilation of citation indexes is especially well suited to the use of powerful computers and does not require indexers who are subject specialists. This helps to make citation indexes more current than most subject indexes.

2. Traditional indexes show only a limited ability to cut across disciplines to pull together related information.

The reason citation indexes do provide multidisciplinary searching capabilities is, once again, related to the fact that most indexers are not as qualified as the author to decide which previously published material is related to his or her current work. A citation index takes advantage of the built-in linkages between documents provided by authors' references by listing together all items with common references. So it identifies relationships between documents that are often overlooked in a subject index. As such, a citation index is an essential tool in informetric work of a multidisciplinary nature.

3. Citation indexes avoid semantic difficulties in preparation and use.

They resolve semantic problems associated with scientific and technical obsolescence by using citation symbols rather than words to describe the contents of a document. So, if for instance, a special technique or scientific law is later called after the first scientist who described it (as in Bradford's law or Peter's principle), the original paper is not indexed by this name (at least not as a subject). However, citation indexes link the original paper to all other papers which cited, and probably used, this technique or scientific law. So, the potential meaning of an item in a citation index is virtually unlimited.

III.2. CITATIONS AND CITERS' MOTIVATIONS

III.2.1. The problem of citers' motivations

Citation is part of the formal process of science. It is used to estimate the quality, impact, originality, penetration or visibility of individual and corporate performance within and across disciplines.

The launching of the SCI in the international scientific community was, however, accompanied by a flurry of correspondence in journals such as Nature and Science, with concern and cynicism (e.g. Cleverdon (1964)) being mixed with cautious interest and welcome (Goudsmit (1974)). It was as if the scientific establishment had not previously recognised the full importance of the citation practice.

Most criticism dealt with the problem of the unknown motivations of citers. Some scientists, such as Polanyi (1966), even maintain that citation is not really a part of the intellectual process, but is rather a skill like cycling or swimming.

There are many reasons why authors cite the works of others. Weinstock (1971) has identified fifteen specific functions of references. These are listed below.

1 Paying homage to pioneers.
2 Giving credit for related work.
3 Identifying methodology, equipment, etc.
4 Providing background reading.
5 Correcting one's own work.
6 Correcting the work of others.
7 Criticising the work of others.
8 Substantiating claims.
9 Alerting researchers to forthcoming work.
10 Providing leads to poorly disseminated, poorly indexed, or uncited work.
11 Authenticating data and classes of fact-physical constants, etc.
12 Identifying original publications in which an idea or concept was discussed.
13 Identifying the original publication describing an eponymic concept or term such as, e.g. Hodgkin's disease ...
14 Disclaiming work or ideas of others.
15 Disputing priority claims of others.

The reasons enumerated in this list might be called 'serious' reasons. Thorne (1977), on the other hand, drew up a list to uncover the citation and publication strategy of some authors. This list can be regarded as a kind of obverse to Weinstock's.

Thorne's list

1 Serial publication (division of a single research project into many parts, each reported separately in LPU's, i.e. least publishable units).
2 Multiple publications (minor variations of a project report submitted to different journals).
3 Hat-tipping citations (acknowledgements of eminent figures).
4 Over-detailed citations.
5 Over-elaborate reporting.
6 Evidentiary validity (citations can be selected to support any point of view).
7 Self-serving citations.
8 Deliberate premeditation (conscious playing of the citation game).
9 Searching out grant funding (identifying currently popular research trends).
10 Funding support for publications (the publication of luxurious research reports to attract attention).
11 Editorial preferences (authors seek to identify preferred topics and styles of journals to which they submit).
12 Citations as projective behaviour (citations as reflection of author biases).
13 Conspirational cross-referencing (the 'you scratch my back and I'll scratch yours' syndrome applied to citation).
14 Pandering to pressures (citing works because it is felt that the reading public requires, or expects, them to be cited).
15 Editorial publication policies (discriminatory biases in editorial policies with respect to selection and rejection).
16 Non-recognition of new authors.
17 Intra-professional feuding.
18 Obsolete citations.
19 Political considerations (citing the 'party line').

It should be noted that some items on this list, such as 2, have more to do with publication counts than with citation counts.

One of the earliest attempts to study citers' motivations was the Moravcsik and Murugesan (1975) paper. They proposed a classification consisting

of eight paired categories. A citation can belong to more than one of the four groups, but not to both categories in any one group. These four groups are :

1 conceptual or operational;
2 organic or perfunctory;
3 evolutionary or juxtapositional;
4 confirmative or negational.

The first of these groups is more of a content than a relationship indicator. It specifies what was cited : a theory, concept or idea (conceptual) or a tool, method or technique (operational). Instead of operational, the term 'methodological' is also used. Group 2 distinguishes between essential and non-essential citations. 'Evolutionary' means that the citing paper builds on previous ideas, while 'juxtapositional' means that an alternative viewpoint is proposed. Group 4 focuses on the citing paper's view about the correctness of the cited work.

These authors also introduced the concept of redundancy. This means that there are several references to papers, all of which make more or less the same point.

The main result drawn from their sample of papers on high energy physics was that 41 % of the citations fell in the perfunctory category.

The work of Chubin and Moitra (1975), often co-cited with that of Moravcsik and Murugesan, was a more or less direct response to their work. Although they recognised the value of an approach to citation analysis based on an inspection of the contents and quality, they recommended a mutually exclusive classification scheme. This alternative, leading to a tree structure, is shown in Fig.III.2.1.

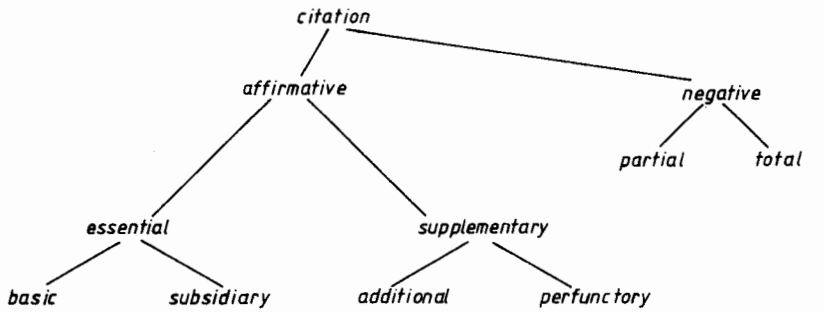

Fig.III.2.1 The Chubin-Moitra classification scheme

In this scheme affirmative citations are either essential or supplementary. Essential citations are further categorised as basic if the

referenced paper is declared to be central to the reported research, a reference on which its findings depend. If a specific method, tool or mathematical result is not directly connected to the subject of the paper, but is still essential to the reported research, it is categorised as a subsidiary citation. Supplementary citations provide additional information when the referenced paper contains an independent supportive observation (idea or finding) with which the citer agrees. Included in the perfunctory category are papers referred to as related to the reported research without additional comment. Finally, the division of negative citations into partial and total needs no further comment.

They found only 20 % perfunctory citations and about 5 % negative citations, none of which was totally negational.

So, in view of these many and divergent approaches it is certainly not surprising to see how authors such as Cronin (1981b, 1984) decried the lack of detailed knowledge concerning the phenomenology of citation and required a 'theory of citing'. Cronin (1981b) suggested that a direct questioning of authors would be the most sensible method of approach. This simple suggestion has been taken up by T. Brooks (1985), whose findings we will discuss in the next section.

III.2.2. An investigation of citers' motivations : the work of Terrence Brooks (1985)

III.2.2.1. Motivational scales

In this study each selected faculty member (26 in total) was asked to assess his or her motivations for giving each reference in one of his or her papers, according to seven scales. Each scale was calibrated from zero, indicating that the scale was not relevant, to three, indicating that the scale was a primary motivation. These seven motivational scales were as follows :

1 Currency (up-to-dateness) scale
Line and Sandison (1974; p.313) discussed the prestige factor driving authors to show how up-to-date they are by referring only to the latest publications.

2 Negative credit
Authors have been described as criticising, correcting, disclaiming and disputing other works by giving negative references. Cole and Cole (1971) defended the view that work of little value will not be criticised (it will hardly be read) but, on the other hand, a highly

criticised paper may be a fruitful error that makes a significant contribution to the field. Items 5, 6, 7 and 14 in Weinstock's list clearly fall within this category.

3 Operational information
Moravcsik and Murugesan (1975) defined an operational reference (cf. above) as one in which a concept or a theory is referred to as a 'tool'. Operational references are made when mathematical or physical techniques or results are borrowed from the cited paper. This coincides with 'identifying methodology and equipment', item 3 in Weinstock's list.

4 Persuasiveness
Gilbert (1977) regards scientific papers as tools of persuasion. In his paper he describes the need of scientists to convince their peers of the correctness of their methods and results.

5 Positive credit
Several items from Weinstock's list such as paying homage to pioneers, giving credit to related work and so on, can be considered as positive credit.

6 Reader alert
References may alert the reader to new, different or obscure sources (item 10 in Weinstock's list), provide background reading (item 4) or identify original publications (items 12 and 13).

7 Social consensus
Bavelas (1978) argued against the evaluative use of citation analysis because she believed that many references are made because of no other reason than an unspecified and vague perception of a consensus in the field (cf. item 14 in Thorne's list).

III.2.2.2. Results

A chi-square test of association on the seven citer motivations proved that there was a significant ($p < 0.005$) difference between these possible motivations. More than half of the observed chi-squared value was being contributed by persuasiveness, which had received many more threes and much fewer zeros than expected. Negative credit, on the other hand, had received many more zeros and much fewer threes than expected. As a whole, three groups could be distinguished : persuasiveness had the highest mean score; positive credit, currency, reader alert and operational information formed the second group; and social consensus and negative credit formed the third group with the lowest mean scores. There was, however, a marked difference between

authors from the exact sciences and those from the humanities, giving evidence that citer motivation may differ by subject area.

III.2.2.3. Further investigations

Another important investigation into citers' motivations was carried out by Vinkler (1987). He characterised the strength of cognitive pressure to cite a paper. The lowest value of cognitive pressure resulting in a citation is called the 'citation threshold'. According to Vinkler, this citation threshold depends primarily on the professional relevance of the potentially citable paper with respect to the study of the citing author. Vinkler's work has been the subject of a comment by Garfield (1989a).

III.2.3. Assumptions underlying citation analysis and problems concerning the use of citation data

III.2.3.1. Basic assumptions

In this subsection we will describe the basic assumptions underlying the use of simple citation counts (Smith (1981)). Some of these clearly conflict with some of the citers' motivations as described in the above section. We note four important assumptions for citation analysis.

1 Citation of a document implies use of that document by the citing author.

According to Smith (1981), this assumption actually has two parts :

a) The author refers to all, or at least to the most important, documents used in the preparation of his or her work.

b) All documents listed were used, i.e. the author refers to a document only if that document has contributed to his or her work.

Failure to meet this criterion for good citation behaviour leads to 'sins of omission and commission'. It is, however, evident that what is cited is only a small percentage of what is read. For this point we also refer the reader to Jones (1976), where fuzzy set theory is used to describe the influence of published documents on an author's work. Moreover, a document is often only cited because of the use of a small part of it, for instance one particular theorem or result.

2 Citation reflects the merit (quality, significance, impact) of that document.

The underlying assumption in the use of citation counts as quality indicators is that there is a high positive correlation between the number of citations which particular document receives and the quality of that document. Of

course, if most citations were made on the basis of Thorne's list, conclusions stemming from citation analysis could easily be invalidated. Moreover, quality and impact are different notions and in practical applications a careful distinction has to be made. Nevertheless, this second assumption has been tested and has found support in a number of studies (see Section III.2.5).

3 Citations are made to the best possible works.

If one assumes that citations are made to the best possible works, then one must imagine that authors sift through all of the possible documents that could be cited and carefully select those judged to be the best. But studies of references behaviour have suggested that accessibility is probably as important as quality as a factor in the selection of an information source. For instance, Soper (1976) found that the largest proportion of documents cited in authors' recent papers was located in personal collections, a smaller proportion was located in libraries in departments and institutions to which respondents belonged, and the smallest proportion was located in libraries in other cities and countries.

4 A cited document is related in content to the citing document.

The fact that citation indexes can be used to retrieve relevant documents supports this assumption.

Other more refined assumptions could be made, some of which will be considered when studying bibliographic coupling and cocitation analysis. We think, however, that these four assumptions form the basis for all studies based on citation counts.

III.2.3.2. Objections to the use of citations

Given the difficulties with the assumptions which underly citation analysis, it comes as no surprise that people raise strong objections to the use of citation counts. Moreover, there are also problems with the sources of citation data. We will now give a short survey of both categories of objections, based on the lists published by Smith (1981) and Vinkler (1986).

1 *Self-citations*

When self-citations are to be eliminated from citation counts, this can easily be done for papers written by a single author. But multi-authored papers may require further checking. Even the definition of the word 'self-citation' is not clear. This problem will be considered in more detail in the next section.

2 *Multiple authorship*

Cited articles listed in the citation indexes include only the first-named authors. To find all citations to publications of a given author, including those in which he or she is not the first author, one needs a bibliography of

his or her works so that all articles can be checked in the citation index. For publications in one of the source journals of ISI's data base, an online search can be of great help here.

There is also the problem of allocating credit in multi-authored works. Should such works be treated the same as single-authored works in citation counts or should credit be divided proportionally? We will also study this problem in the next section.

3 *Homographs*

Many scientists with the same name and initials could be publishing in the same field. To differentiate among them additional information such as institutional affiliation is needed. Otherwise citations could be attributed incorrectly. This problem is even greater with Chinese or Japanese names than with English ones (Cornell (1982)).

4 *Synonyms*

Citations will be scattered unless a standard form for the author's name can be established. Examples of such 'synonyms' include an author's name with a variable number of initials, a woman's maiden and married names, different transliterations of foreign (e.g. Russian or Chinese) names, and misspellings. A famous example in the field of informetrics is given by the different forms of Derek J. de Solla Price's name.

Journal names may also create problems here. In addition to variations in the abbreviated form of a given title, journals merge, split into new journals, change titles and appear in translations (Garfield (1975)).

5 *Types of sources*

The types of sources used in citation analysis can influence the results, as demonstrated in a study by Line (1979) in the social sciences. Analyses of references drawn from journals and monographs showed differences, some of them large, in data distributions, forms of materials cited, subject self-citation, citations beyond the social sciences, and countries of publication cited. So, the choice of types and numbers of sources should depend on the purpose of the analysis.

6 *Implicit citations*

Most citation analyses consider only explicit citations, for the simple reason that the SCI and the SSCI only give information on this kind of citation. As explained in Section III.1.4, the A&HCI includes some implicit citations too. Implicit citations are also frequently found in the form of eponyms. By an 'eponym' we mean an expression that consists of an individual's name plus a word denoting some idea or thing associated with that person. Examples in informetrics include : Bradford's law of scattering, Lotka's inverse square

law, Zipf's law and so on. Furthermore, papers containing important ideas will not necessarily continue to be highly cited. Once an idea is sufficiently widely known, citing the original version is unnecessary (e.g. Einstein's theory of special relativity). This phenomenon has been termed 'obliteration by incorporation'. In the case of citation analysis we should also mention many instances of implicit references to Eugene Garfield as editor of the SCI, the SSCI and the A&HCI. Most people feel free to use these tools without having to mention Garfield or his collaborators.

7 *Fluctuations in time*

There may be large variations in citation counts from one year to another, so that citation data should not be too restricted in time. An attempt to account for random fluctuations of journal citations was made by Nieuwenhuysen and Rousseau (1988).

8 *Field variations*

Publication performance, i.e. the number of publications and publication practices, depends strongly on specialities. As a consequence citation counts also strongly depend on research fields. This leads to difficulties in cross-discipline comparisons (Pinski and Narin (1976; p.298)).

9 *The incompleteness of the ISI-database*

There are numerous periodicals which are not covered by the ISI-database. This might create problems for local studies (Velho (1986, 1987), Gaillard (1989)).

10 *Dominance of English as a scientific language*

The English language dominates the scientific community, especially in the Western world. As a consequence papers published in English are preferred for citations (Vinkler (1986), Garfield (1979b)). This means that if two papers make the same point but one of them is written in English and the other one in a different language (French, Russian, Dutch, ...), the one written in English would get significantly more citations. Moreover, this bias seems to be greater in the social sciences than in the exact sciences.

11 *The 'American' bias*

Especially the A&HCI but also, though to a lesser extent, the Science and Social Sciences files of the ISI-database show a bias towards publications from the USA. Moreover, it has been found (Cronin (1981a)) that on the average, US authors cited American works for 95 % (much more than the US's share of the world's scientific output) and British authors for roughly 40 %. Inhaber and Alvo (1978) found that the US literature attracted by far the greatest number of citations. US journals were approximately seventeen times as likely to cite themselves or other US journals than journals from the UK. UK journals on the other hand divided their attention almost equally between the US and

the UK literature.

12 *Sex bias*

As shown by Marianne Ferber (1986), in some cases researchers tend to cite a larger proportion of authors of their own sex than of the opposite sex. This has substantial consequences in fields where men constitute a large majority.

13 *Errors*

Of course, citation analyses, including those based on citation indexes, can be no more accurate than the raw material used. The incorrect citing of sources is unfortunately far from uncommon as shown, for instance, by Goodrich and Roland (1977). Besides these errors, which might be called random errors, there are also 'systematic' errors, such as the underestimation of citations since, for example, preprints can only be indexed under 'in press' or 'unpublished'.

In this section we have considered two types of limitations which can affect citation analyses : the assumptions made may not be true or the collected data may have inadequacies. Invalid conclusions will be made unless these limitations are taken into account when designing a study and interpreting results. The most reliable results may be expected when citation abuses and errors appear as noise, where this noise represents only a relatively small number of the citations analysed.

III.2.4. Self-citations and co-authorship

In our survey of problems surrounding the use of citation counts we already mentioned 'self-citations' and 'multiple authorship'. Here we wish to discuss these notions in more depth.

III.2.4.1. Self-citations

The term 'self-citation' has been used with different meanings. If the citing paper has one or more authors in common with the cited paper one usually describes this feature as self-citation. However, references to articles published in the same journal in which the citing article appears are also said to be self-citations. This kind of self-citation will be discussed further in Chapter III.5. When citations are used for science policy purposes (see Chapter III.7) citations of articles authored by people working in the same scientific institution or in the same research group as the citing author are also called 'self-citations' (see, for example, Moed et al. (1985a,b) and Vinkler (1986)). Earle and Vickery (1969) used the term 'self-citation' to indicate a similarity of subject matter (the same scientific field or subfield) between citing and cited article, as opposed to 'self-derivation'

or the citing of papers from other topics or scientific areas. To distinguish between the first-mentioned one and these other types of self-citation, the former are sometimes called 'author self-citations'. It is this type which will be discussed in this section.

Tagliacozzo (1977) conducted a systematic study of author self-citations in the areas of plant physiology and neurobiology. We reproduce here the main conclusions of her study.

There are very few articles which do not include any self-citation and the numbers of self-citations per article are widely distributed. Moreover, authors are inclined to cite their own work more abundantly than the work of any other single author. Self-citations refer to a more recent group of publications than other citations do. This recentness of self-citations might indicate the high degree of continuity in the work of individual scientists.

One of the important findings of Tagliacozzo is that more self-citations than other citations are repeatedly cited in research articles. If we assume that frequency of repetition of the same citation in the text of an article is an indication of the significance of the cited work for the citing work (see also Rousseau (1987b)), then the set of self-citations clearly stands out as having a particular prominence among the other references of the citation network.

Contrary to expectation, the results of the study showed no significant relationship between the size of the bibliography and the extent of self-citing. There was also no significant relationship between the amount of self-citing and the productivity of the author.

Meadows (1974; pp.159-161) suggests that the self-citation proportion may reflect the 'maturity' of a specialty. Applying this view to the individual, the number and proportion of non-self-citations might then indicate the maturation of an individual's research (see also Porter (1977)).

Lawani (1982) distinguishes two genera of author self-citations : synchronous self-citations and diachronous self-citations. An author's synchronous self-citations are those contained in the references the author gives (those studied by Tagliacozzo), whereas diachronous self-citations are those included in the citations an author receives.

For example, if an author writes a paper with 12 references, 3 of which are to this author's own work, the synchronous self-citation rate (a better term would be self-citing rate) would therefore be 25 %. In general, an author's self-citing rate over a certain period is determined by considering all the papers he or she has published or co-authored in this period, finding the number of the author's own papers listed in the references, and expressing

this as a percentage of the total number of references in all the papers. To determine an author's diachronous self-citation rate (self-cited rate), one notes how many times the works of this author are cited in the period under study and how many of these citations were made by papers in which the author's name appeared. One then calculates what percentage is constituted by these self-citations.

Lawani (1982) indicates the practical implications of high or low self-citing and self-cited rates. For instance : high self-cited rate certainly implies egotism on the part of the author. On the other hand, a researcher's self-citing rate may be high but if this researcher is also cited heavily by others, the self-cited rate would be low. Such a researcher is not an egotist. Indeed, a relatively high self-citing rate coupled with a low self-cited rate may well suggest that the researcher concerned is a productive and key figure in his or her research specialty.

Reports on the number of self-references indicate for science subfields an average of 5 to 20 %. Individual authors may, however, have a much greater amount of self-references. Porter (1977) finds that it does not matter whether one includes self-citations, at least when using citation analysis to study science subfields. If one wants to use citations to review individual scientists, Lawani's approach may probably prove more useful.

III.2.4.2. Co-authorship

We will now turn to the problem of co-authorship. Citation analysis of collaborative studies is difficult because the ISI citation indexes list citations only by senior (i.e. first) author. Counting citations to junior-authored works requires collaborators to be identified and hence an author search and a search for specific citations under their names. If this is not done, how serious is the resultant loss of information? And, if it is done, how should the contribution of a co-author be weighted?

Generally speaking, there are three ways to deal with the problem of which weight a co-author deserves. These are : straight count, when no account is taken of multiple authorship and the paper is allocated to the first author; adjusted count, which gives every collaborator (or at least some collaborators) some fraction of the authorship and finally, normal count, giving full credit to all contributors (see also Bookstein (1984) and Subsection IV.2.2.2).

J.R. Cole and S. Cole (1973) claim that 'the omission of collaborative citations to papers on which the author was not the first among collaborators does not affect substantive conclusions'. They therefore recommend the use of straight counts. Using straight counts solves the problem of distributing

credit for multiple-authored work, as the first author receives all the credit. Moreover, it greatly reduces the work required to collect the data. Lindsey (1980) notes that the straight count approach can be considered as a sampling strategy. As such, it should be examined in terms of its representativeness.

The straight count procedure presupposes that the set of papers on which a scientist's name occurs first (including solo-authored papers) is a representative sample of all of that scientist's papers. To study this, we first remark that it would be reasonable to assume that the name order of authors listed on a given paper reflects the level of their contributions, with the greatest contributor listed first, and so on in descending order. However, when the authors are listed alphabetically, this usually is not the same as the order of contribution. A comprehensive examination of 1500 chemists conducted by Rudd (1977) found a greater percentage of first authors among those with last names beginning with A to F compared to G to M, and with G to M compared to N to Z (the percentages of first authors in the three groups were 56.8, 29.9 and 13.3 respectively). Lindsey (1978) also found that in a stratified random sample of publications in seven fields, there was a significantly greater probability that, among members of a collaborative team, the member whose name occurs first alphabetically also appeared first in the list of authors. We should further note that Zuckerman (1968) and Garfield (1982) found that eminent scientists received nearly twice as many citations as secondary authors than as primary ones. This shows that eminent scientists have been systematically ceding primary authorship to their junior collaborators. This practice was termed 'noblesse oblige' by Harriet Zuckerman (1968). Along with Lindsey (1980) we may conclude that there is neither strong empirical evidence nor theoretical rationale to support the underlying assumptions for using straight counts.

This leads us to the problem of how to attribute papers or citations to papers to the different contributors. Normal count gives full credit to all contributors. But, as studied by Lindsey (1980), this procedure tends to inflate the publication or citation scores of those who produce many multi-authored papers. There is also a small technical problem here : the sum of the number of publications of every author is not equal; in fact, strictly speaking, it is larger than the number of papers under study.

So, finally, the best way to handle multi-authored papers is to assign credit proportionally. Thus, if a paper is written by two authors, each would get half a credit. Three authors would get each a third, and so on (as proposed, for example, by Lindsey (1980) and Price (1981a)). This method is called 'adjusted counts'.

Along with Lindsey (1980), Long et al. (1980) and Garfield (1979a, p.243) we wish to conclude that the only fair way of developing relative citation counts is to compile the performance of all the published material that is listed in a comprehensive bibliography and to use adjusted counts.

III.2.5. In support of citation analysis

The use of citation analysis as an indicator of scholarly merit is surrounded with controversy. One of the reasons might perhaps be that it opens the door to the evaluation of scientific research by outsiders. Here we will review some of the papers written in favour of citation analysis. We are of the opinion, however, that the subsequent chapters of this part, illustrating different uses of citation analysis, constitute the main argument in favour of this technique as a valuable informetric method.

Probably the best paper that compares peer judgement with citation-based judgement was written by Julie Virgo (1977). Her study had two objectives. The first was to test the hypothesis that in a given discipline journal articles that are cited more frequently will tend to be judged more important than articles that are cited less frequently. The second was to identify by means of a regression analysis other factors associated with articles that were judged important.

In Virgo's experiment a group of medical researchers was asked to evaluate the importance of articles. To obtain the articles for the study, a bibliography tailored to each of the participant's specific research or clinical interests was drafted. The participant then evaluated each item in the bibliography in order to rate it according to its relevance to his or her own research interest. Only those articles judged 'very relevant' were retained.

After the lists of relevant articles were compiled, the Science Citation Index was used to find all papers which cited them. Only citations made during the first three years from the date of publication were counted. Once citation frequencies were collected for all relevant articles in a participant's bibliography, the articles were ranked according to the frequency with which they had been cited. The top five and the bottom five articles in the ranking were selected, then pairs of articles were formed on the basis of one member selected at random from the frequently cited group and one member at random from the infrequently cited group. Five pairs of articles were then submitted to each participant together with a series of questions to which he or she was asked to respond with reference to each pair.

Each participant was further asked to name two persons anywhere in the United States whom he or she considered to be outstanding persons working on

the same research as the participant. One of these two persons was asked if he or she would be willing to participate in the study and was then sent the same set of articles as the first participant, together with a questionnaire about the importance of the papers.

It was found that the association of citation frequency and importance judgements was statistically significant on the 5 % level. Moreover, citation frequency on the average predicts the more important paper better than the second judge did. These results make a strong case in favour of citation analysis.

As to the second purpose of Virgo's study, it was found that the impact factor of the journal in which the paper was published contributed significantly, together with citation frequency, towards explaining the variability of ratings. (The impact factor of a journal is the average number of citations received by each paper published during a specified time period, see Chapter III.5.)

Citations often strongly correlate to other measures of eminence. In a book on the profession of psychology in the US, Clark (1957, Chapter 3) reported on a study in which a group of 'highly visible' psychologists was selected, and then each of them was asked to choose twenty-five persons who 'had made the most significant contributions to psychology as a science' either because of published work or through the training of doctoral candidates. The final list correlated with other measures of eminence, such as the extent of publications, offices held in the American Psychological Association or listings in biographical directories. He found moreover that citation counts correlated the highest (Pearson's R = 0.67) with status as a psychologist. This showed that citation analysis can do at least as well as other measures to measure eminence.

Subsequent work by Myers (1970) also provides a strong argument for the validity of citation analysis. He compared lists of the most frequently cited authors in psychology with fifteen independent measures of eminence and found that citation frequency is a good index of a scientist's esteem. We should also mention Narin (1976), who reviewed 24 studies showing that citation counts, as well as other informetric measures, correlate in the range of R = 0.5 to R = 0.8 with various rankings of eminence.

Studies have shown that scholars rely heavily on citations to locate library materials. One of the most thorough investigations in this direction has been conducted by Broadus (1977). In this extensive review he came to the conclusion that despite some inconsistencies, there appeared to be a strong relationship between citation counts and other methods of evaluating science. Moreover, in a more recent review, Bensman (1982) found a growing consensus

about the validity of citation analysis.

We conclude this section by stating (cf. Garfield (1983)), that citation analysis is not a substitute or shortcut for critical thinking; it is, instead, a point of departure for those willing to explore the avenues to thorough evaluation. Although peer review and citation analyses are highly correlated, there is enough variance to warrant using both procedures in tandem.

III.2.6. Notes and comments

Thorne's list addresses the issue of the ethics of citation practices. Garfield has commented on this on several occasions (e.g. Garfield (1982)). He stresses the fact that to cite someone is to acknowledge that person's impact on subsequent work. As such, citations form an important part of the reward system in science. Failure to cite one's sources, which Garfield considers a form of plagiarism, robs the individual of the credit he or she deserves. Outright plagiarism is rather exceptional in science (but see, for example, Illinois (1985)). Citation amnesia, on the other hand, is a common phenomenon. 'Citation amnesia' refers to a researcher using an idea or concept he has seen or heard about somewhere, without crediting the original source.

Many people find that the use of eponyms should be discouraged. Henwood and Rival (1980) found the use of eponyms to be 'symptomatic of shoddy work and conducive to trivial work', because the eponym is only descriptive and does not have any inherent meaning. Using eponyms is naming things after people, rather than scientific content.

Concerning differences in citation rate due to the nature of the article itself or the field about which the paper deals we mention the following. Peritz (1983) found that in sociological journals methodological papers are more frequently cited than theoretical or empirical ones. Moreover, she found that this result was not due to a few outliers (very highly cited papers). Here Peritz defined methodological papers as papers dealing with methods of study design, data collection and analysis. Theoretical papers are papers that discuss concepts and general theoretical schemata, and empirical papers are any investigations which use empirical data of whatever source.

Field differences are illustrated by the fact that the average chemistry or physics article contains about twenty references, while mathematical articles contain less than ten. The fewer references in maths papers may say something about the literary style of mathematicians who write more esoterically than other scientists (Garfield (1976b)). It should be mentioned also that, for instance in mathematics, some subjects are studied by so few people, that, however brilliant a paper might be, it will nevertheless receive

only a few citations (this has to do with the problem of superspecialisation).

The difference in citation behaviour between American journals and journals which publish mainly papers written in a language other than English is also illustrated by Wellisch (1980), who makes a comparison between the Journal of the American Society for Information Science and Nachrichten für Dokumentation. Among many differences we note that JASIS contained only 2 % of references to documents not written in English, but that Nachrichten für Dokumentation contained almost 29 % references to documents written in a language other than German.

Many journals from developing countries are de facto not available to scientists in the developed countries. Hence these journals are not cited and are not included in the ISI database (Velho (1986, 1987)). The 'invisibility' of Third World science has also been studied by Gaillard (1989), who shows that 'low quality' is certainly not the only reason to explain this situation. To increase the number of Third World scientific journals which are included in the ISI database a cooperative program has been launched under the name of 'The Philadelphia Program' (Moravcsik (1988a)). It appears, however, that hardly any of the proposed measures have been converted into reality so far and hence the Program is subject of severe criticism, not in the least coming from developing countries. For the special case of Indian scientific journals, this problem has also been studied by Manorama and Bhutiani (1988).

III.3. CITATION NETWORKS AND CITATION MATRICES

III.3.1. Generalities on citation graphs and citation matrices

When a document d_i cites a document d_j, we can show this by an arrow going from the node representing d_i to the node representing d_2 (Fig.III.3.1). In this way the documents from a collection D form a directed graph, which is called a '*citation graph*' or '*citation network*'. The latter term is often used for the particular case in which the graph is connected. In the incidence matrix of this graph (called the 'citation matrix') a 1 is placed in each cell in which the row corresponds to the citing document d_i and the column to the cited document d_j. Table III.3.1 illustrates this.

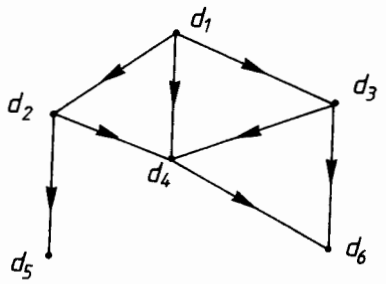

Fig.III.3.1 Citation net

Table III.3.1. The incidence matrix corresponding to Fig.III.3.1

	d_1	d_2	d_3	d_4	d_5	d_6
d_1	0	1	1	1	0	0
d_2	0	0	0	1	1	0
d_3	0	0	0	1	0	1
d_4	0	0	0	0	0	1
d_5	0	0	0	0	0	0
d_6	0	0	0	0	0	0

Of course, it is also possible to reverse the arrows to obtain the graph of the relation 'is cited by'. The incidence matrix of this graph is merely the transpose of the incidence matrix of the citation network of the relation 'cites'.

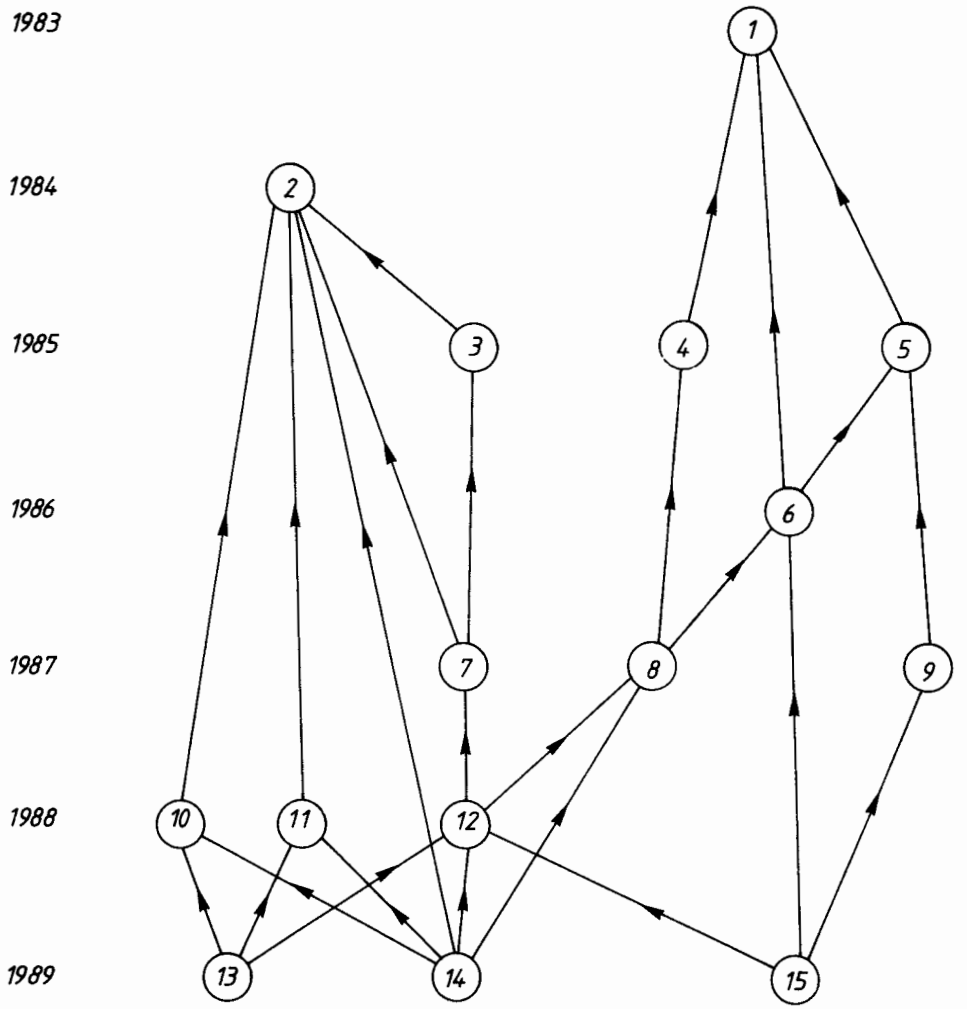

Fig.III.3.2 Citation net where the vertical axis denotes time

In Fig.III.3.2 circles representing articles are arranged in horizontal rows by year of publication, with the most recent year at the bottom. As it is a citation graph, an arrow means 'cites'.

From this graph several deductions can be made without any knowledge about the subject content (Cawkell (1974)). Paper number 2 has had a considerable impact upon later work, since it has been heavily cited. Papers

13 and 14 are probably rather similar in subject content as they contain common references to articles 10, 11 and 12 : the former are bibliographically coupled by the latter (for more details on the notion of bibliographic coupling see Section III.4.1). If they had contained more references in common, their subject content would have a still greater degree of similarity. Until 1988, the articles in Fig.III.3.2 formed two disconnected groups. In that year 7 and 8 were co-cited by 12 (the notion of co-citation will be studied in Sections III.4.2 and III.4.3). The link between the two groups was consolidated the next year as, for instance, 8 and 12 were co-cited by 14, and 6 and 13 by 15. This implies that the relatedness between the two groups was first perceived by the author of 12.

III.3.2. Some mathematical theorems on citation graphs

Proposition 1 (Kochen (1974, p.17)).
Let d be any document and let C(d) be the set of all references in d; similarly, let $C^{-1}(d)$ be the set of all documents from which d received a citation. If d_0 is now a fixed document, then

$$d_0 \in \bigcap_{d \in C(d_0)} C^{-1}(d) \; . \qquad [III.3.1]$$

The statement of this proposition is obvious once we have understood the mathematical symbolism. Formula [III.3.1] merely states that when we form the collection of all documents cited by d_0 - this is $C(d_0)$ - and we pick any d in this collection, then d_0 belongs to the set of all documents that cite d. See also Fig.III.3.3.

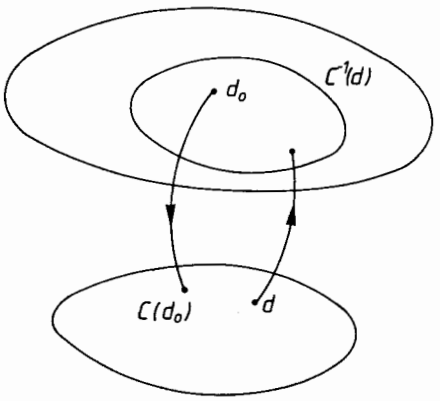

Fig.III.3.3 Illustration of Proposition 1

Corollary (Egghe) :
For every document d_0 we have :

$$\bigcap_{d \in C(d_0)} C^{-1}(d)$$

is the set of all documents d_1 such that the reference list of d_1 includes the reference list of d_0.

Proof. To show equality between two sets we have to prove two inclusions. Let d_1 first belong to $\bigcap_{d \in C(d_0)} C^{-1}(d)$, then d_1 belongs to $C^{-1}(d)$, and this for every d in $C(d_0)$. This means that every d in the reference list of d_0 also belongs to the reference list of d_1. This proves the first inclusion.

Suppose next that the reference list of d_1 includes the reference list of d_0. Then $C(d_0) \subset C(d_1)$ and hence $\bigcap_{d \in C(d_1)} C^{-1}(d) \subset \bigcap_{d' \in C(d_0)} C^{-1}(d')$. Now, by Proposition 1, $d_1 \in \bigcap_{d \in C(d_1)} C^{-1}(d)$, which proves the other inclusion. □

The next result is somewhat more intricate. First, we recall that a directed graph is said to be weakly connected if for any two vertices there is a path of edges joining them, where the directions at the edges are ignored.

Theorem 2 (based on Kochen (1974, p.18)).
If the citation graph of a non-empty, finite set D of n documents is weakly connected, then, for any d_0 of D :

$$D = \bigcup_{j=0}^{N-1} C_j \quad , \qquad \text{[III.3.2]}$$

where

$$C_j = \bigcup_{d_j \in C_{j-1}} [C(d_j) \cup C^{-1}(d_j)] \quad , \quad j > 0$$

and

$$C_0 = \{d_0\} \ .$$

Proof. Pick any $d_0 \in D$, where $C(d_0)$ is the set of all references of d_0 and $C^{-1}(d_0)$ is the set of all citations of d . Then C_1 is the set of all documents which either cite d_0 or are cited by d_0. By the requirement of weak connectedness, C_1 is not empty unless D is equal to the singleton $\{d_0\}$, in which case the

result of the theorem is trivial. So we can proceed and form C_2. By Proposition 1 we know that $d_0 \in C_2$.

To show that $\bigcup_{j=0}^{N-1} C_j$ is equal to D, we suppose that some $d \in D$ does not belong to $\bigcup_{j=0}^{N-1} C_j$. However, this assumption leads to a contradiction : as there is a path, necessarily finite, joining d to d_0, there is a number $j \leq N-1$ such that $d \in C_j$. □

This theorem yields an algorithm for obtaining all the documents of a given collection, provided the collection is reasonably homogeneous, so that its citation graph is weakly connected. Moreover, if D is a large computer file, then the algorithm gives a procedure for exploring the core of a topic (take d_0 to be a core document) and moving further and further towards the boundaries. This method is known as 'cycling'.

The following result gives useful insight into the structure of a citation matrix, cf. Kochen (1974, p.21).

Theorem 3 :
The average number of references per document times the number of documents being considered is equal to the average number of citations to a fixed document times the total number of different references.

Proof. Let C be the citation matrix of the collection under study. Therefore $c_{ij} = 1$ if document d_i cites reference r_j and $c_{ij} = 0$ if it does not; the columns contain only those documents that are cited at least once by the documents in the collection. Now, if there are n source documents, the average number of references per document will be

$$\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} c_{ij} \right) .$$

Here p is the total number of different references and $\sum_{j=1}^{p} c_{ij}$ is the number of references contained in d_i. Then the average number of references per document times the number of documents is equal to

$$\sum_{i=1}^{n} \sum_{j=1}^{p} c_{ij} ,$$

which is the total number of 1's in the citation matrix. On the other hand, we see that the average number of citations to a fixed document is

$$\frac{1}{p} \sum_{j=1}^{p} \left(\sum_{i=1}^{n} c_{ij} \right) ,$$

where $\sum_{i=1}^{n} c_{ij}$ is the number of citations to r_j (the j^{th} reference). Then, the average number of citations times the total number of different references is also equal to the total number of 1's in the citation matrix. This proves the theorem. □

An application. Suppose you have a (20,120) citation matrix and you know that the average number of references per document is 10. Then the preceding result states that the average number of citations equals (10)(20)/120 = 1.67.

III.3.3. The publication and citation process described by matrices

The aim of this section is to show how often-used notions such as 'number of co-authors' or 'number of articles published in one year' can be described in purely mathematical terms.

Following Krauze and McGinnis (1979), we assume that for a given field of knowledge and a fixed period T, a matrix W is given. This (m,n)-matrix $W = (w_{ij})$ is defined by

$$w_{ij} = \begin{cases} 1 & \text{if author i contributed to paper j ,} \\ 0 & \text{otherwise .} \end{cases} \qquad \text{[III.3.3]}$$

The fact that W is an (m,n) matrix implies that we consider m authors and n papers. Let U be the column vector in $\mathbb{R}^n$ consisting of 1's only. Hence $U = (1,1,\ldots,1)^t$. Then we can use matrix terminology to describe the following concepts :

1 The number of contributions of author i (in the period T) given as :

$$\sum_{j=1}^{n} w_{ij} = (WU)_i = (WW^t)_{ii} \quad . \qquad \text{[III.3.4]}$$

2 The number of co-authors of paper j :

$$\sum_{i=1}^{n} w_{ij} = (U^tW)_j = (W^tW)_{jj} \quad . \qquad \text{[III.3.5]}$$

3 The number of papers to which both authors i and j contributed (number of collaborations) :

$$\sum_{k=1}^{n} w_{ik}w_{jk} = (WW^t)_{ij} \quad . \qquad \text{[III.3.6]}$$

4 The number of authors shared by papers i and j (number of common contributors) is :

$$\sum_{k=1}^{m} w_{ki} w_{kj} = (W^t W)_{ij} \quad . \qquad \text{[III.3.7]}$$

Let us now consider m source papers that cite n different references and their citation matrix $C = (c_{ij})$. In analogy with the notions described above we obtain the following :

5 The number of references of a given paper d_i is :

$$\sum_{j=1}^{n} c_{ij} = (CU)_i = (CC^t)_{ii} \quad . \qquad \text{[III.3.8]}$$

6 The number of citations received by a given paper r_j is :

$$\sum_{i=1}^{m} c_{ij} = (U^t C)_j = (C^t C)_{jj} \quad . \qquad \text{[III.3.9]}$$

7 The number of references which d_i and d_j have in common, called the 'bibliography coupling strength' is given by

$$\sum_{k=1}^{n} c_{ik} c_{jk} = (CC^t)_{ij} \quad . \qquad \text{[III.3.10]}$$

8 The number of citations which r_i and r_j have in common (the 'co-citation strength') is given by :

$$\sum_{k=1}^{m} c_{ki} c_{kj} = (C^t C)_{ij} \quad . \qquad \text{[III.3.11]}$$

This shows the universal power of the matricial notation.

III.4. BIBLIOGRAPHIC COUPLING AND CO-CITATION ANALYSIS

III.4.1. Bibliographic coupling

'*Bibliographic coupling*' is a term introduced by M.M. Kessler (1963b) of the Massachusetts Institute of Technology. He defined a unity of coupling between two papers as an item of reference used by these two papers. The two papers are then said to be bibliographically coupled. Their bibliographic coupling strength is then the number of references they have in common (cf. Subsection III.3.3.7). Fig.III.4.1 illustrates this notion. It shows two papers with a bibliographic coupling strength of 3.

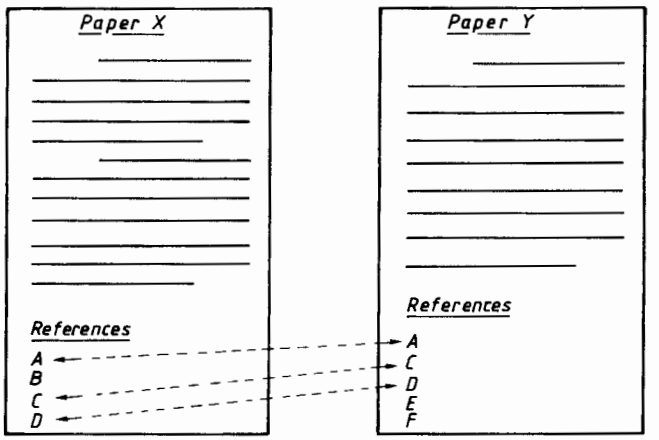

Fig.III.4.1 Two papers, bibliographically coupled with a strength equal to three

Kessler (1963a) defined two criteria of coupling, criterion A and criterion B.

Criterion A. A number of papers constitute a related group $G_A(P_0)$ if each member of the group has at least one coupling unit in common with a fixed paper P_0. The coupling strength between P_0 and any member of $G_A(P_0)$ is the number of coupling units (n) between them. Then $G_A(P_0;n)$ denotes that subset of $G_A(P_0)$ that is linked to P_0 through exactly n coupling units (cf. Fig. III.4.2).

Contrary to Kessler's original definition, we will state that P_0 is bibliographically coupled to itself, with a coupling strength equal to the number of references in P_0. If, however, P_0 contains no references, it is not bibliographically coupled to itself.

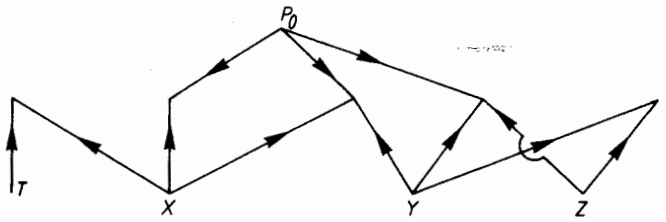

Fig.III.4.2 $\longrightarrow$ means 'cites'
$G_A(P_0) = \{P_0,X,Y,Z\}$
$G_A(P_0;2) = \{X,Y\}$

Criterion B. A number of papers constitute a related group G_B if each member of the group has at least one coupling unit to every other member of the group.

The subset $\{P_0,X,Y\}$ in Fig.III.4.2 is an example of a G_B; Z cannot be included in this subset as it has no coupling unit with X. As a test, Kessler (1963a) treated each article (265 in total) in volume 97 of Physical Review as a P_0. Thus 265 G_A's were generated. The number of articles in each G_A varied from one (not coupled to any of the other articles) to twelve. This experiment proved the existence of the G_A-phenomenon in a population of well-edited papers in a well-established field of science.

We will now give some simple formulae concerning G_A, $G_A(n)$ and G_B. Most of the easy proofs are left to the reader; see also Rousseau (1987c).

(F1) If a paper has a non-empty reference list, then the set consisting of this single paper is a G_B; $P_0 \in G_A(P_0)$ for every P_0.

(F2) If $P_1 \in G_A(P_0)$, then $P_0 \in G_A(P_1)$.

(F3) If $P_1 \in G_A(P_0)$ and $P_2 \in G_A(P_0)$, then we can generally only conclude that $P_1 \in G_A(P_2)$ if P_0 contains just one reference.
Proof. If P_0 contains more than one reference (e.g. R_1 and R_2), then the situation in Fig.III.4.3 is possible.
If however, P_0 contains exactly one item of reference, say R_1, then necessarily $P_1 \in G_A(P_2)$, and vice versa, by (F2).

(F4) $\bigcup_{n \in \mathbb{N}_0} G_A(P_0;n) = G_A(P_0)$.

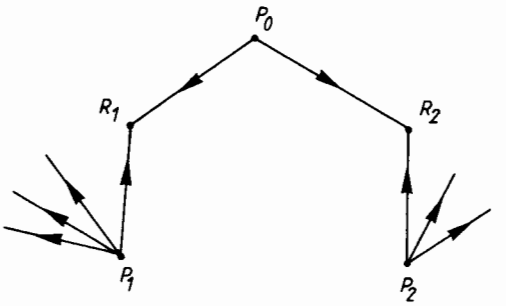

Fig.III.4.3 $P_1 \notin G_A(P_2)$

(F5) If $m \neq n$; $m,n \in \mathbb{N}_0$, then

$$G_A(P_0;n) \cap G_A(P_0;m) = \emptyset \quad .$$

(F6) If $P_1 \in G_B$ and $P_2 \in G_B$, then $P_1 \in G_A(P_2)$ and vice versa.

(F7) If we denote the set of citing papers under consideration by Ω and if

$$\forall P_0 \neq P_1 \in \Omega : P_1 \notin G_A(P_0) \quad ,$$

then either P_1 has no references
or P_1 has only unique references (i.e. not contained in any other item in Ω).
Note that the expression '$\forall P_0 \in \Omega : P_1 \notin G_A(P_0)$' implies that $P_1 \notin G_A(P_1)$ and hence that P_1 contains no references.

(F8) If $P_0 \in G_B \subset \Omega$, then

$$\begin{array}{ccccccccc} & & \bigcap\limits_{P_1 \in \Omega} G_A(P_1) & & & & & & \\ & \subset & & \cap & & & & & \\ \emptyset & & & & \bigcap\limits_{P_1 \in G_B} G_A(P_1) \subset G_A(P_0) \subset \bigcup\limits_{P_1 \in G_B} G_A(P_1) \subset \bigcup\limits_{P_1 \in \Omega} G_A(P_1) \subset \Omega & & & & \\ & \cap & & & & & & & \\ & G_B & \subset & & & & & & \end{array}$$

Proof. We only prove the least obvious inclusions.

(i) $G_B \subset \bigcap_{P_1 \in G_B} G_A(P_1)$, for if $Q \in G_B$ and $P_1 \in G_B$ then by the definition of G_B, Q and P_1 have at least one reference in common and hence $Q \in G_A(P_1)$. As this is true for every $P_1 \in G_B$, we see that $G_B \subset \bigcap_{P_1 \in G_B} G_A(P_1)$.

(ii) $\bigcap_{P_1 \in G_B} G_A(P_1) \subset G_A(P_0)$.

If $Q \in \bigcap_{P_1 \in G_B} G_A(P_1)$, then $Q \in G_A(P_0)$ as $P_0 \in G_B$.

(iii) If $P_0 \notin \bigcap_{P_1 \in \Omega} G_A(P_1)$, i.e. if there is at least one $P_1 \in \Omega$ such that P_0 has no reference in common with P_1, and if we take $G_B = \{P_0\}$, then we have an example in which $\bigcap_{P_1 \in \Omega} G_A(P_1)$ and G_B cannot be compared.

(F9) If $\# \Omega = m$, then $\forall P_0 \in \Omega$, $\forall i > m : G_A(P_0;i) = \emptyset$.

(F10) $\# G_A(P_0) = \sum_{n=1}^{\# \Omega} (\# G_A(P_0;n))$.

(F11) In a set of papers, each containing at least one reference, the relation 'is bibliographically coupled with' is reflexive, symmetric but not transitive.

Kessler saw bibliographic coupling in the first place as a retrieval tool. Knowing a given paper P_0 to be relevant to a user's research, the retrieval system would also retrieve $G_A(P_0)$, i.e. all papers bibliographically coupled with P_0. As a retrieval tool, bibliographic coupling has the following properties :

1 Bibliographic coupling is independent of words and language. All the processing is done in terms of numbers. All difficulties of language, syntax and word habits are thus avoided. This is an advantage bibliographic coupling shares with all citation-based techniques (cf. Section III.1.5).
2 No expert reading or judgement is required. Indeed, the text need not even be available. Again this advantage is shared with all citation-based techniques.
3 The group of papers associated with a given test paper extends into the past as well as the future. As a paper continues to be cited, the group of papers bibliographically coupled with it also grows.
4 The method does not produce a static classification for a given paper. The groupings will undergo changes that reflect the current usages and

interests of the scientific community.

Finally, Kessler (1963a) remarks that a paper's G_A could be considered as its 'logical references'.

According to Bella Weinberg (1974), bibliographic coupling should work best (in the sense that the results show little bias) for review papers because they should need to cite a lot of older papers. So, bibliographic coupling should tend to favour repetitive literature.

The major theoretical criticism of the concept of bibliographic coupling comes from Martyn (1964). He contends that a bibliographic coupling unit is not a valid measure of relationship because the fact that two papers have a reference in common is no guarantee that both papers are referring to the same piece of information.

A critical review of bibliographic coupling was written by B. Weinberg (1974), to which the reader is referred for further information. In recent times co-citation analysis, a variant of bibliographic coupling - to be discussed in the following sections - has become more popular. However, the new SCI Compact Disc edition employs bibliographic coupling to show the user those papers which are most closely related to the retrieved paper. In this compact disc version twenty such related papers can be traced in the order of their coupling strength.

III.4.2. Co-citation : part I

A Soviet information scientist (Irina Marshakova) and an American one (Henry Small) independently proposed the same variation on bibliographic coupling. Small (1973) and Marshakova (1973) both suggest using *co-citation* of documents as a method of measuring relationships between documents.

Two documents are said to be co-cited when they both appear in the reference list of a third document. The co-citation frequency is defined as the frequency with which two documents are cited together. Thus, while bibliographic coupling focuses on groups of papers which cite a source document, co-citation focuses on references which frequently come in pairs (see Fig.III.4.4). In the Soviet literature bibliographic coupling is said to be 'retrospective' and co-citation is called 'prospective coupling' (Marshakova (1973)).

Alternatively, by using the language of set theory, we can define the co-citation frequency of two documents X and Y as follows. If A is the set of papers that cite document X and B is the set of papers that cite document Y, then $A \cap B$ is the set of documents that cite both X and Y. The number of elements in $A \cap B$, denoted as $\#(A \cap B)$, is the co-citation frequency of X and Y.

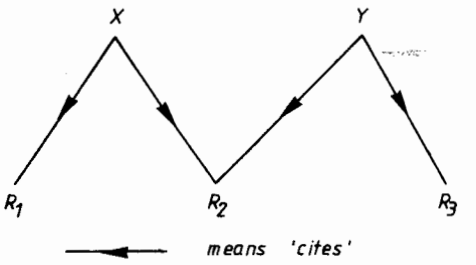

Fig.III.4.4 X and Y are bibliographically coupled; R_1 and R_2 are co-cited, as are R_2 and R_3

Of course, to be precise one should also mention the period under consideration. It makes quite a difference whether one studies co-citation over a one year or a ten year period.

The relative co-citation frequency can easily be defined in terms of set theory. Using the above notation, it is written as

$$\frac{\#(A \cap B)}{\#(A \cup B)} \quad . \qquad [III.4.1]$$

(Note that one can similarly also define the notion of relative bibliographic coupling strength).

In practice, the co-citation frequency of two scientific papers can be determined by comparing lists of citing documents in a citation index (or its online or on disc version) and counting identical entries. Recall that in Section III.3.3 we defined co-citation in terms of citation matrices.

In a way similar to bibliographic coupling, one can define two co-citation criteria.

Criterion A. Several papers constitute a co-citation related group M_A (or better $M_A(P_0)$) if each member of the group is co-cited at least once with a given paper P_0. $M_A(P_0;n)$ then denotes that subset of $M_A(P_0)$ consisting of those papers that are exactly n times co-cited with P_0.

Criterion B. Several papers constitute a co-citation related group M_B if each member of the group is co-cited (at least once) with every other member of the group.

It is now easy to state and prove results analogous to (F1)-(F11) for the co-citation relation (Rousseau (1987c)).

Small (1973) saw co-citation analysis mainly as a way to map out in great detail the relationships between the key ideas in a scientific field, leading to a more objective way of modelling the intellectual structure of scientific specialties. He illustrated these ideas by the co-citation network for that specialty of particle physics which might be as described as 'Theories of broken chiral symmetry and current algebras' (cf. Fig.III.4.5 and Table III.4.1).

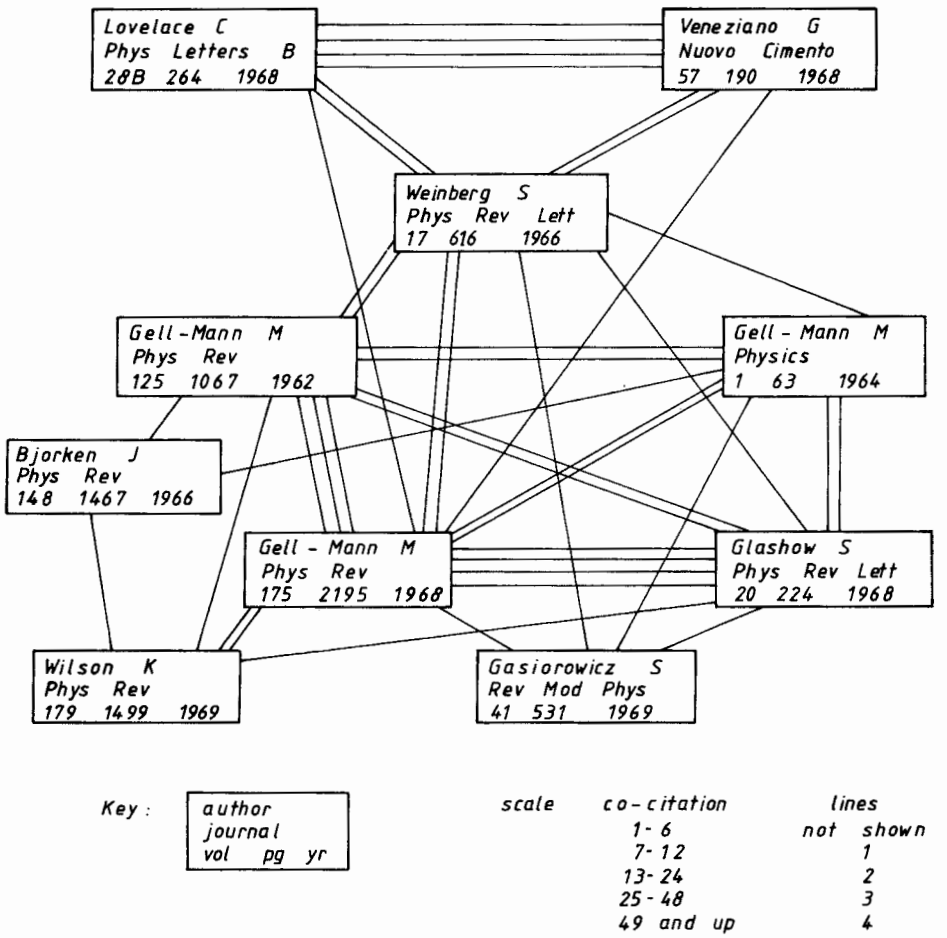

Fig.III.4.5 Co-citation network for frequently cited papers in particle physics (data from the 1971 SCI; graph redrawn from Small (1973))

Table III.4.1. Papers in the network in Fig.III.4.5

Bjorken J.D., Applications of the chiral U(6)⊗U(6) algebra of current densities. Physical Review 148, 1467, 1966.

Gasiowicz S. and Geffen D.A., Effective Lagrangians and field algebras with chiral symmetry. Reviews of Modern Physics 11, 531, 1969.

Gell-Man M., Symmetries of baryons and mesons. Physical Review 125, 1067, 1962.

Gell-Mann M., The symmetry group of vector and axial vector currents. Physics 1, 63, 1964.

Gell-Mann M., Oakes R.J. and Renner B., Behavior of current divergences under $SU_3 x SU_3$. Physical Review 175, 2195, 1968.

Glashow S.L. and Weinberg S., Breaking chiral symmetry. Physical Review Letters 20, 224, 1968.

Lovelace C., A novel application of Regge trajectories. Physics Letters B, 264, 1968.

Veneziano G., Construction of a crossing-symmetric, Regge-behaved amplitude for linearly rising trajectories. Nuovo Cimento, 57, 190, 1968.

Weinberg S., Pion scattering lengths. Physical Review Letters, 17, 616, 1966.

Wilson K.G., Non-Lagrangian models of current algebra. Physical Review, 179, 1499, 1969.

This network shows two pairs of heavily co-cited papers from 1968 : Lovelace - Veneziano and Gell-Mann - Glashow, connected through the earlier work of Gell-Mann and Weinberg.

Co-citation patterns are found to differ significantly from bibliographic coupling nets, but to agree generally with patterns of direct citation (Small (1973)). Generally speaking, an interpretation of the significance of strong co-citation links must rely both on the notion of subject similarity and on the association or co-occurrence of ideas. As such, co-citation can also be used in information retrieval. For instance, a secondary index based on highly cited papers would allow sequential searches through a citation index, retrieving a list of new documents at each co-cited entry point. Co-citation could also be used to establish a cluster or core for a particular specialty This idea has been realised in ISI's Atlas of Science (Garfield (1981, 1987)). Recently, Sharabchiev (1989) compared clusters obtained by bibliographic coupling and by co-citation analysis.

Martyn's (1964) criticism on bibliographic coupling also applies to co-citation analysis : the fact that two papers are co-cited does not imply

that they contain similar pieces of information. The strongest opponent of co-citation analysis without human judgement is probably David Edge. In his two papers 'Why I am not a co-citationist' (1977) and 'Quantitative measures of communication in science : a critical review' (1979), he expresses his view that quantitative methods such as co-citation analysis have only a limited use. Among other objections he emphasises the need to be able to see individual variations. It is often because individual scientists and groups do not share the consensus view, as shown by co-citation maps, that crucial innovative decisions are made. Another critical evaluation of co-citation analysis was written by Sullivan et al. (1977).

III.4.3. Co-citation : part II

In this section we will discuss *tri-citation* and the *circle model* to represent this and other forms of multiple citation. Further, we will propose a tentative classification of couples of papers, according to their relative co-citation. Finally, we will investigate the statistical validity of co-citation graphs as a function of co-citation strength for a given value of citation frequency.

III.4.3.1. The circle model for multiple citation patterns (Small (1974))

In Section III.4.2 we have defined the co-citation frequency of two documents X and Y in terms of set theory, namely as the number of elements in the intersection of the set of papers citing X and the set of papers citing Y. Similarly, we can define the notions of tri-citation and n-citation as the number of elements in the intersection of 3 or n sets of citing documents. These notions then lead to higher levels of aggregation, tri-citation being the next higher level.

The central idea of the circle model is that a document may be represented as a circle in which the area is proportional to the number of papers which cite it, i.e. the citation frequency is set equal to $k\pi r^2$, with k as a constant of proportionality and r as the radius of the circle. Hence knowing the citation frequency yields r. This model leads to the following interpretation of co-citation and tri-citation frequency. If two documents are co-cited, the frequency of co-citation will be equal to k times the area of intersection of the two circular areas; if three documents are tri-cited, the frequency of tri-citation will equal k times the area of tri-citation of the three circles. Since for a group of three documents the radii are completely determined by the citation frequencies and the distances between their centres by the co-citation frequencies, the area of tri-section has been

determined, and this area may be used as a predicted value for the real tri-citation count.

Small (1974) has tested this for the 1972 citation data of six documents in particular physics (see Table III.4.2).

Table III.4.2. Citation frequencies of six papers (1972 data)

Document	Code	citation frequency
D. Amati, A. Stanghellini and S. Fubini, Nuovo Cimento, 26, 896-954, 1962	A	88
J. Benecke, T.T. Chou, C.N. Yang and E. Yen, Phys. Rev., 188, 2159-2169, 1969	B	127
L. Caneschi and A. Pignotti, Phys. Rev. Lett., 22, 1219-1223, 1969	C	39
C.E. DeTar, C.E. Jones, F.E. Low, J.H. Weis, J.E. Young and C.I. Tan, Phys. Rev. Lett., 26, 675-676, 1971	D	76
R.P. Feynman, Phys. Rev. Lett., 23, 1415-1417, 1969	F	187
A.H. Mueller, Phys. Rev. D, 12, 2963-2968, 1970	M	142

As all of the papers were frequently cited in 1972 and all were co-cited with each other, it was possible to calculate all 15 distances among the six documents, using the equation given below and the co-citation frequencies of any two documents X and Y as above.

The number of co-citations of X and Y :

= the area of intersection of their citation circles,

= the area of the sector XZFS + the area of the sector YZES - the area of the quadrilateral XZYS,

$$= \frac{\hat{ZXS}}{2} r_X^2 + \frac{\hat{ZYS}}{2} r_Y^2 - d_{XY}.r_X.\sin\left(\frac{\hat{ZXS}}{2}\right) \qquad \text{[III.4.2]}$$

(for the meaning of the symbols see Fig.III.4.6).

Moreover, we have :

$$\frac{\hat{ZXS}}{2} = \arccos\left(\frac{r_X^2 - r_Y^2 + d_{XY}^2}{2r_X d_{XY}}\right), \quad \text{(in radians)} \qquad \text{[III.4.3]}$$

$$\frac{\hat{ZYS}}{2} = \arccos\left(\frac{r_Y^2 - r_X^2 + d_{XY}^2}{2r_Y d_{XY}}\right) \quad \text{(in radians)} \qquad \text{[III.4.4]}$$

and

$$\sin\left(\frac{\hat{ZXS}}{2}\right) = \left[1 - \left(\frac{r_X^2 - r_Y^2 + d_{XY}^2}{2r_X d_{XY}}\right)^2\right]^{1/2} , \qquad \text{[III.4.5]}$$

where d_{XY} denotes the distance between X and Y. Since r_X and r_Y are determined from the citation counts, d_{XY} can be found from the co-citation counts and the above equations. For practical cases we also need an iterative computer procedure, as d_{XY} cannot be written in an explicit form.

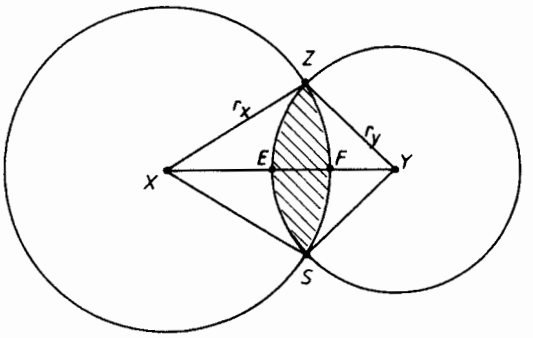

Fig.III.4.6 Two-citation circles and their area of intersection

This gave the following matrix of distances (Table III.4.3) :

Table III.4.3. Distance table for the papers of Table III.4.2

	A	B	C	D	F	M
A	-	8.15	6.51	7.34	8.81	8.08
B		-	5.09	7.72	4.40	6.50
C			-	4.69	6.38	6.05
D				-	8.58	5.35
F					-	8.19
M						-

Note that we do not have to fill in the lower half of this matrix as distances are symmetric. Using this matrix, a slight variation of Kruskal's multidimensional scaling method (Kruskal (1964); see also Section I.5.3) was used to obtain the configuration shown by Fig.III.4.7. Once this configuration was found it was possible to predict the tri-citation frequencies and to compare them with the actual observed frequencies. As there are six documents,

one has to consider $\binom{6}{3} = 20$ groups of three documents (see Table III.4.4). A chi-square test, with 20 degrees of freedom, gave agreement on the 10 % level (critical value : 28.41).

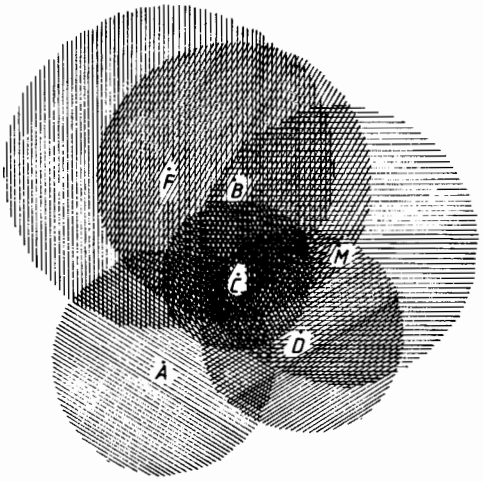

Fig.III.4.7 Circle model for the data of Table III.4.2 (redrawn from Small (1974))

Table III.4.4. Circle model : predicted and observed tri-citation counts

	tri-citation	observed	predicted	chi-square
1	BFM	41	39.81	0.036
2	BCF	24	22.60	0.087
3	BDF	19	16.25	0.465
4	ABF	18	19.57	0.126
5	DFM	18	18.62	0.021
6	CFM	17	16.31	0.029
7	BCM	16	16.94	0.052
8	BDM	15	17.07	0.251
9	CDF	14	11.97	0.344
10	AFM	13	12.13	0.062
11	BCD	13	11.27	0.266
12	CDM	11	13.65	0.514
13	ABM	9	11.37	0.494
14	ADM	8	9.72	0.304
15	ACF	7	6.96	0.000
16	ABC	7	6.60	0.024
17	ACM	6	5.86	0.003
18	ADF	5	6.31	0.272
19	ABD	3	4.94	0.762
20	ACD	2	4.56	1.437
				5.549

The success of the circle model in predicting tri-citation counts strongly supports the interpretation of a cited document as an area in subject space, rather than as a geometrical point.

III.4.3.2. Some types of co-citation histories

To measure to what extent two documents are considered to be related, one uses the relative co-citation. Indeed, after some initial build up, the number of citations to a given paper usually decreases. Hence the co-citation frequency of two papers also declines. However, when more and more scientists consider two papers as being strongly related, their relative co-citation will increase. In the limit the relative co-citation, which is always between 0 and 1, can attain the value of 1. An example of such an occurrence is given in Table III.4.5.

Table III.4.5. Citation and co-citation history of two papers between 1966 and 1982 (data from the SCI; compiled by R. Rousseau). Convergence of relative co-citation to 1.

Y : year

K : citations to : 'Kawarabayashi K. and Suzuki M., Partially conserved axial-vector current and the decays of vector mesons. Phys. Rev. Lett., 16, 255-257, 1966'

R : citations to : 'Riazuddin and Fayyazuddin, Algebra of current components and decay widths of ρ and K* mesons. Phys. Rev., 147, 1071-1073, 1966'

C : co-citation frequency of K and R

RC : relative co-citation of K and R (calculated from [III.4.1]).

Y	K	R	C	RC
1966	21	6	3	0.13
1967	51	31	25	0.44
1968	82	67	58	0.64
1969	54	41	36	0.61
1970	39	38	31	0.67
1971	31	24	22	0.67
1972	18	14	11	0.52
1973	23	18	18	0.78
1974	18	11	11	0.61
1975	11	4	3	0.25
1976	6	5	5	0.83
1977	7	7	6	0.75
1978	6	5	5	0.83
1979	4	6	4	0.67
1980	4	4	4	1.00
1981	7	4	4	0.57
1982	3	3	3	1.00

A case such as illustrated in Table III.4.5 is very rare. Somewhat more frequently one has (in the notation of Section III.4.2) :

$$\max \left\{ \frac{\#(A \cap B)}{\#A}, \frac{\#(A \cap B)}{\#B} \right\}$$

in short : MAX(X,Y), tends to 1. This happens, for instance, when the content of one paper coincides with a part of the content of another. However, this kind of convergence can also happen when one of the two papers is written by a very prestigious author and the other paper is written by a lesser known scientist : people citing the lesser known also cite the better known but not vice versa. An example of this kind of convergence is shown in Table III.4.6.

Table III.4.6. Citation and co-citation history of two papers between 1972 and 1982 (data from the SCI, compiled by R. Rousseau). Convergence of MAX to 1

Y : year

T : citations to : ' 't Hooft G. and Veltman M., Regularization and renormalization of gauge fields. Nucl. Phys. B, 44, 189-213, 1972'

A : citations to : 'Ashmore J.F., A method of gauge-invariant regularization. Nuovo Cimento Lett., 4, 289-290, 1972'

C : co-citation frequencies of T and A

MAX : MAX(T,A)

Y	T	A	C	MAX
1972	5	0	0	-
1973	24	9	5	0.56
1974	52	23	14	0.61
1975	46	9	3	0.33
1976	46	11	7	0.64
1977	42	9	6	0.67
1978	23	5	1	0.20
1979	48	10	7	0.70
1980	89	9	6	0.67
1981	80	10	10	1.00
1982	69	10	10	1.00

Of course, more frequently, one finds that two papers are co-cited at some higher or lower level without any convergence to one. These papers are somewhat related but also deal with unrelated topics. An example of this situation is shown in Table III.4.7.

Table III.4.7. Citation and co-citation history of two papers between 1975 and 1982 (data from the SCI, compiled by R. Rousseau). Case of relatively low co-citation frequencies

Y : year

D : citations to : 'De Rujula A., Georgi H. and Glashow S.L., Hadron masses in a gauge theory. Phys. Rev. D, 12, 147-162, 1975'

A : citations to : 'Augustin J.E. et al (35 co-authors), Discovery of a narrow resonance in e^+e^- annihilation. Phys. Rev. Lett., 33, 1406-1408, 1974

C : co-citation frequencies of D and A

RC : relative co-citation of D and A

Y	D	A	C	RC
1975	6	279	4	0.01
1976	42	183	11	0.05
1977	101	110	19	0.10
1978	97	51	8	0.06
1979	71	46	7	0.06
1980	67	33	9	0.10
1981	104	26	13	0.11
1982	78	25	4	0.04

In some rare cases the relative co-citation even decreases. A possible explanation of this phenomenon is that one of the authors (or group of authors) of the two papers has written an improved version, containing new facts, about the same subject. Table III.4.8 gives an example of two papers with decreasing relative co-citation frequencies. However, in this particular case, we do not know the exact reason for the decline.

Table III.4.8. Citation and co-citation history of two papers between 1974 and 1982 (data from the SCI, compiled by R. Rousseau). Decreasing relative co-citation frequencies

Y : year

M : citations to : 'Moncada S., Ferreira S.H. and Vane J.R., Prostaglandins, aspirin-like drugs and the oedema of inflammation. Nature, 246, 217-219, 1973'

W : citations to : 'Williams T.J. and Morley J., Prostaglandins as potentiators of increased vascular permeability in inflammation. Nature, 246, 215-217, 1973'

C : co-citation frequencies of M and W

RC : relative co-citations of M and W

MAX : MAX(M,W)

Y	M	W	C	RC	MAX
1974	9	11	8	0.67	0.89
1975	17	11	8	0.40	0.73
1976	20	22	14	0.50	0.70
1977	26	33	18	0.44	0.69
1978	18	25	11	0.34	0.61
1979	19	17	9	0.33	0.53
1980	14	26	7	0.21	0.50
1981	9	19	5	0.22	0.56
1982	12	27	7	0.22	0.58

Finally, to end this overview of co-citation histories we remind the reader of the fact that the vast majority of papers pairs is never co-cited at all!

III.4.3.3. Statistical validity of co-citation graphs (Shaw (1985))

In a co-citation graph, points denote documents and two points are joined when the co-citation strength of the associated pair of documents is equal to or greater than some threshold value. In many cases, the use of a co-citation threshold yields a disconnected graph in which documents are distributed among the components. In certain investigations such as those performed by Small and Griffith (1974), these components are interpreted as scientific specialties and subspecialties.

To investigate the value of these partitionings in components and hence of the interpretations derived from them, Shaw (1985) has tested the so-called Random Graph Hypothesis. This means that one assumes that the lines of a co-citation graph (at a certain threshold value) are randomly selected from the set of all possible lines. This assumption constitutes the null hypothesis which defines a graph for which there is no meaningful clustering structure

(this issue has also been mentioned in Subsection I.5.4.5).

The results of his tests suggest that as the co-citation threshold increases, meaningful pairwise associations are broken, related documents appear in different components and the partition becomes statistically invalid. Also, as the co-citation threshold decreases, meaningless pairwise associations are created, unrelated documents appear in the same component and again the partition becomes statistically invalid. Consequently, this suggests that there may exist two critical values of the co-citation threshold which define the limits of statistical validity. Between these thresholds, the results are statistically valid and can be interpreted. Outside the region of validity, the results can be attributed to the clustering technique and not to the existence of an inherent structure in the data.

III.4.4. Citation context analysis

In this section we will discuss the work of citation analysers who have focused on the citation context, hence taking less interest in the type of citing-cited relationship. Here we should remark that Moravcsik's and Murugesan's first category (conceptual/operational) is actually a content classifier.

Garfield (1970) and Small (1978) argued that cited documents become in some sense symbols for the ideas they contain. In this view referencing becomes a labelling process. Moreover, this is also one of the explanations of the utility of citation indexes.

To explore how cited documents behave as 'standard symbols', Small (1978) studied some highly cited documents in chemistry. He showed, first, that individual citation contexts may be regarded as instances of symbol deployment; second, he determined to which extent this symbolic content is shared among the citing authors. To do this, he defined the notion of uniformity of usage, which is the percentage of citing contexts which share the prevalent view on the cited item. Table III.4.9 shows some of his results. The table reveals the great uniformity with which these documents are cited : the main uniformity of usage for books was 68 percent, and was 92 percent for journal articles. The fact that books have a smaller uniformity of usage is certainly not very surprising in view of their broader information content.

Table III.4.9. Citation uniformity of usage (Small (1978))
J : journal article; B : book or monograph

cited item	concept	percent uniformity
Stewart (J)	hydrogen scattering factors	93
Lowry (J)	protein determination	100
Cromer (J)	atomic scattering factors	93
Pauling (B)	atomic radii	56
Woodward (J)	orbital symmetry rules	72
Pople (B)	CNDO and INDO methods	100

Complete references :

Stewart R.F., Davidson E.R. and Simpson W.T., Coherent x-ray scattering for the hydrogen atom in the hydrogen molecule. Journal of Chemical Physics, 42, 3175, 1965.

Lowry O.H., Rosebrough N.J., Farr A.L. and Randall, R.J., Protein measurement with the Folin phenol reagent. Journal of Biological Chemistry, 193, 265, 1951.

Cromer D.T. and Waber J.T., Scattering factors computed from relativistic Dirac-Slater wave functions. Acta Crystallographica, 18, 104, 1965.

Pauling L., The nature of the chemical bond (Cornell University Press, Ithaca) 1960.

Woodward R.B. and Hoffmann R., The conservation of orbital symmetry. Angewandte Chemie, 81, 797, 1969.

Pople J.A. and Beveridge D.L., Approximate molecular orbital theory (McGraw-Hill, New York) 1970.

Another observation made by Small (1978) was the high frequency with which these works were involved in so-called redundant citations. Redundant citations can be taken to indicate one of a number of situations. For example, they may signal simultaneous and independent discovery. Another reason for redundant citations might be the availability of several good sources for the same concept or procedure.

Highly cited documents can be considered as 'exemplars' or 'concept symbols' : illustrations of methods or theories which comprise the essential repertoire of techniques for practitioners of a specialty. Some papers then occupy distinct niches in an author's cognitive space, representing the concept they are associated with. Treating references as part of the symbol system of language links citation and co-citation analysis to that part of the psychology literature that studies associative recall, word recognition and artificial intelligence. Because references and ideas in text appear in conjunction with other references and ideas (co-citation), this yields a

complex system, which is, to some extent, shared by other authors and which is continually changing (Small (1987)). Specialty narratives incorporated in ISI's Atlas of Science are a practical result of citation context analysis (Small (1986)).

One implication of this view is the possibility that the certified contents of a document can have different meanings for different research groups and can change or shift in time. Some cases of these phenomena have been studied by Small (1985) and by Cozzens (1982, 1985).