THREE DREXEL INFORMATION SCIENCE RESEARCH STUDIES

BY

RALPH GARNER
LOIS LUNIN
LOIS BAKER

EDITED WITH AN INTRODUCTION BY

BARBARA FLOOD

DREXEL PRESS 1967

CONTENTS

Chapters

CONTENTS (cont'd)

Tables

Illustrations

CONTENTS (cont'd)

CONTENTS (cont'd)

Tables

Illustrations

CONTENTS (cont'd)

Illustrations (cont'd)

Chapters

CONTENTS (cont'd)

Tables

Illustrations

CONTENTS (cont'd)

Illustrations (cont'd)

INTRODUCTION

Three Drexel Information Science student research studies are reproduced in this volume of the Drexel Library School Series. The student authors are the following:

> Ralph Garner, who received the Drexel M. S. in Information Science in June 1965, is now President of Ralph Garner Associates, New York, and a doctoral student at New York University. Garner also has a B. S. in Physics from the City College of New York.
>
> Lois Lunin, who received the Drexel M. S. in Information Science in June 1966, is now Program Director, Information Center for Human Communication, Welsh Medical Library, Johns Hopkins University, Baltimore. Mrs. Lunin received her B. A. in biology from Radcliffe.
>
> Lois Baker, who received the Drexel M. S. in Information Science in June 1966, is now with the Institute for Scientific Information, Philadelphia. Mrs. Baker also has a B. S. in chemistry from Western Reserve University.

These students are three of the first four graduates from the Drexel Information Science curriculum. The first graduate, Santina Isabella received her M. S. in Information Science in June 1964. Miss Isabella is now with the National Library of Medicine, Bethesda. Her research report was summarized in the Proceedings of the American Documentation Institute (Parameters of Information Science) held in Philadelphia in 1964.

This volume is offered with the hope that full publication will provide essential data for the serious student and that the economies of the paper-back form will bring it into reach of all. Few studies can be clearly and completely understood without full text and data.

A full curriculum in the information sciences was first offered by Drexel in the Spring of 1963. The decision to offer a full curriculum was reached as a result of the enthusiastic response to course offerings in documentation and information science since 1959. These course offerings were, in turn, a logical outgrowth of a generation of emphasis on special librarianship at Drexel. The Information Science program is quite separate from the Library Science curriculum. Most students in the program offer at least a bachelor's degree in a science or technology subject field for admission; many have more advanced work, both in academic and job experiences. A total of 125 students have been admitted to the Information Science curriculum since it began. The curriculum was established to educate indexers, abstractors, translators, literature and patent searchers, computer center staff members, systems analysts and designers, and information center administrators. The program is offered in close cooperation with the Drexel Computer Center, the College of Business Administration and the College of Engineering and Science. The eminence of its part-time faculty has been a distinguishing feature.

Drexel sponsors the Drexel Information Science Series published by the Spartan Press, Washington, D. C. , now in its fourth volume. A list of this Series as well as the Drexel Library School Series is given at the back of this volume with the hope that it will be useful to the library and information science world.

A summary of Mrs. Baker's report is to be published in the Journal of Chemical Documentation. A summary of Mrs. Lunin's report *was* published in Methods of

Information in Medicine and in the Proceedings of the International Federation for Documentation meeting in Washington, 1965.

The following are among the cooperating institutions contributing to these studies: (a) M. D. Anderson Hospital, University of Texas, Houston, (b) Drexel Biology Department, (c) Institute for Scientific Information, Philadelphia, (d) University of Pennsylvania, Moore School of Electrical Engineering, (e) U. S. Air Force, and (f) the Welsh Medical Library, Baltimore.

In carrying out their studies Mr. Garner received assistance from Eugene Garfield, John Harvey, J. B. Maginnis, and Richard Davis; Mrs. Lunin from Isaac Welt, Claire Schultz, Richard Davis and Martin Lunin; and Mrs. Baker from Arthur Elias, John Harvey, Conrad Kruse, and Clarence Van Meter. In addition, the assistance of Antoinette Pasles and Joseph D'Auria in producing this volume should be recognized.

Barbara Flood
Instructor of Information Science
Christmas, 1966

A COMPUTER ORIENTED, GRAPH THEORETIC ANALYSIS OF CITATION INDEX STRUCTURES

by

Ralph Garner

PREFACE

The purpose of this research was to investigate whether or not the mathematical discipline of graph theory is applicable to the analysis of citation indexing, and if it is applicable to identify these areas and perform a graph theory analysis. Graph theory was found applicable. It was possible to analyze all of a list of citation index terms and structures. However, a detailed analysis of the physical representation of the citation index search product has been omitted since it has been found that an analysis of the topological characteristics of a citation index search product requires the application of concepts significantly different from those employed in this report.

The facets of graph theory applicable to citation indexing structures has led to much material which is valuable in socio-historical research. Consequently, this material has been included for completeness.

In order to exhaustively investigate all the material presented above, it is first necessary to develop a mature understanding of all aspects of graph theory. To do this thoroughly would require more time than is required to complete this research. Consequently, the author has attempted to gain an overall knowledge of graph theory and in so doing, has attempted to avoid missing an aspect of graph theory which might be fruitful; but along with this it is admitted that no aspect of the graph theory has been completely or exhaustively utilized. It is with this understanding that the author presents, rather than an exhaustive graph theory analysis, an introduction to a graph theory analysis of citation index structures.

CHAPTER I

INTRODUCTION

The purpose of this research is to investigate whether the mathematical discipline of graph theory is applicable to the analysis of citation indexing, and if it is applicable to identify the areas and perform a graph theory analysis. This paper, therefore, presents a mathematical notation based on that conventionally used in graph theory, and then applies the notation to the many aspects of citation indexing discussed in the literature.

The motivation for this work is the belief that a mathematical notation will permit a clear statement of the problems associated with citation indexing and will consequently facilitate needed solutions. Moreover, it is believed that a mathematical notation can be beneficially used when discussing citation indexing structures so as to permit non-ambiguous communication among researcher, system designer, programmer, and users of citation index data. The notation presented here will afford definitions of the input, output, and processing problems associated with the computer manipulation of citation index data. Additionally, this notation will facilitate a detailed and exact discussion of the use of citation data for socio-historical analysis.

Need for an Analysis of Citation Index Data

The use of printed citation indexes as a tool for retrieving information is well established. At present the use of these printed indexes limits the use of citation indexes to searches which can be completed by a book-form, non-manipulative, index. Most citation index searches are reduced, therefore, to a page-turning procedure, comparable to searches with conventional indexes. It would be desirable to use large-scale, citation index files in machine language for searches involving manipulative techniques of interrogation. In the near future such large-scale citation files will be available.

With the availability of large-scale, citation files on magnetic tape or other media, a host of new problems need solution. Effective searching of large-scale computer files involves major programming tasks. The accomplishment of these tasks is dependent upon one's ability to specify and formulate, in a mathematical notation, both the data and the operations involved. It is necessary, therefore, to unambiguously define the structure of the search question. Only when these two structures are fully understood can one then efficiently proceed to design a powerful system for storage and retrieval of citation data. The problems which each of these structures present is outlined below.

The Structure of Citation Index Data

The use of mathematical descriptions almost inevitably creates a synthesis of what has already been said, along with a succinct and manipulative compressing of ideas. This has not yet happened to the literature on citation indexing. For example, such terms as "connectedness," “cycling, " “critical path,” “bibliographic coupling,” which frequently appear in the literature, have not been sufficiently examined -- mathematical definitions at present do not exist. It is these terms and others like them that can be brought clearly into focus through the use of mathematical notation. The use of such a notation would provide clear-cut descriptions so that problems could be defined for programmers. A mathematical notation, therefore, is needed to describe the structure of citation data so that computer-oriented tasks can be performed, and thereby permit the use of large-scale citation files in an information retrieval system.

Structure of the Search Query

Asking a machine-form citation index system a question presents a number of new problems. In the MEDLARS system the formulation of a query is in the form of a Boolean statement: Give me everything you have on (A) roentgenological techniques in (B) Schmorl's disease in (C) geriatric patients; or in Boolean notation $A \cap B \cap C$. But what kinds of questions can we ask in a citation index system, and more importantly, how do we formulate a question in a citation index system? While it may be true that Boolean symbolism is useful, it certainly is far from a panacea. Moreover, for a citation system, Boolean symbolism alone, is completely inadequate. The problem, therefore, is to provide a computer-oriented procedure for formulating search questions.

Graph Theory

Graph theory is a branch of mathematics which belongs in part to topology and more closely in subject matter and method to combinatorial analysis. It also occurs, sometimes under other names, in electrical network analysis, organic chemistry, theoretical physics and statistical mechanics and the group dynamics aspects of social psychology, Graph theory, loosely defined, treats of graphs wherein pairs of points are joined by lines.

Interest in graph theory and its applications has fluctuated considerably since its early development two hundred years ago. At that time Euler grappled with the celebrated Konigsberg bridge problem. However, it took many years for other applications of graph theory to be found. As early as the late 19th century, Cayley (1821-1895) founded the study of matrices and showed how matrices could be used to represent graphs. Moreover, Cayley was the first to apply graph theory to the problem of representing chemical structures. For several decades potential applications and the value of graph theory went unnoticed. Recently interest in graph theory has developed in the information sciences.

Summary of the Literature

In 1955 Garfield’ first proposed the use of citation indexes for science. Recently he has surveyed the literature of this field and discussed its many applications,[2] and the references he listed have been omitted from the bibliography of this paper. Garfield[3] had also discussed the use of graphs in citation indexing in the social sciences and expanded some earlier notions on citation networks.[4]

Graph theory as an established branch of mathematics has few comprehensive texts.

[1] E. Garfield, "Citation Indexes for Science -- A New Dimension in Documentation Through Association of Ideas, " Science, 122 (3156): 108-111 (1955).

[2] E. Garfield, "Science Citation Index -- A New Dimension In Indexing," Science, 144 (3619): 649-654 (1964).

[3] E. Garfield, "Citation Indexes in Sociological and Historical Research, " American Documentation, 14 (4): 289-291 (1963).

[4] E. Garfield, "Citation Indexing: A Natural Science Literature Retrieval System for the Social Sciences, " The American Behavioral Scientist, 7 (10): 58-61 (1964).

The standard English works are Berge5 and Ore6 which have been available since 1962. The number of articles on the application of graph theory, however, is vast. Flament,[7] for example, in applying graph theoretic notions to group structures has over 85 bibliography items.

This research reports the application of that portion of graph theory which is useful in defining and analyzing a citation network. The papers related to this research are listed in the Bibliography. Of these authors none discuss graph theory in this particular context.

Included in the Bibliography are supporting mathematical texts. This background material is necessary since the ability to apply graph theoretic notions involves such mathematical disciplines as matrix theory, mathematical logic, theory of sets, and abstract algebra.

Proposed Method for Solution

The structures considered above -- the structures of the data and the search query (and their associated problems) -- can be formulated conveniently and succinctly by developing and applying an appropriate system of notation. This notation will be especially useful for describing and manipulating citation structures. It is proposed that the notation conventionally employed in graph theory, combined with matrix theory, can be used to provide efficient, succinct, and useful solutions to this problem. The resulting notation will enable one to state explicitly the various structures necessary for manipulation and retrieval of citation index data. These structures can take the form of a mathematical formula, a graphic representation, and a matrix representation. The mathematical formulation is necessary for the statement of the search query. The graphic representation is necessary for the search product. And the matrix representation is necessary for the computer analysis and manipulation of the citation index file.

By using such a notation to describe each type of structure, we are then in a position to begin the design of a large-scale, computer-oriented citation index information storage and retrieval system.

[5] C. Berge, The Theory of Graphs and Its Applications, John Wiley and Sons, Inc. , New York (1962).

[6] 0. Ore, Theory of Graphs, American Mathematical Society, Providence, Rhode Island (1962).

[7] C. Flament, Applications of Graph Theory to Group Structures, Prentice-Hall, Inc. , Englewood Cliffs, New Jersey (1963).

CHAPTER II

EXPLANATION OF SYMBOLS

The purpose of this section is to indicate very briefly the meaning of the symbols used in this work which are not explained elsewhere.[1] These symbols are common in other areas of algebra as well as in graph theory.

An ordered set is a group or collection of items where each item is unique. In the equation, S = (xi, $x_2, x_3\}$, S is the set which has as its members $x_1, x_2,$ and x3. The braces are used to enclose the members of the set. When defining a set it is sometimes to our advantage to use a delta under the equals sign, $\underset{\Delta}{=}$, to discriminate between an equality and a definition. The "roof" symbol) , appears only in an equation which defines a set. The roof is placed over the symbol which will be used when referring to the members of the defined set in general. This symbol is immediately followed by a statement in braces about the characteristics of the members of the set. Thuss $\underset{\Delta}{=} \hat{x}$ (x is a scientific paper) is an equation which says that the set S is defined as the elements x, where x is the symbol to be used when referring to these elements in general; and each x has the property that it is a scientific paper.

If it is necessary to indicate the number of members in a set, this is accomplished by enclosing the set by two vertical lines. The equation, $|S| = 3$ says that the number of members of the set S is equal to 3. The epsilon ε , is used to indicate set membership. Therefore, $y \varepsilon X$ says that y is a member of the set X. The symbol $\wedge$ is read "restricted to." The Boolean operation $\cup$ and $\cap$ maintain the conventional meaning of union and intersection respectively. The term partition is used to mean the decomposition of a set into subsets which are mutually exclusive and jointly exhaustive.

We shall adopt the convention that when a symbol x (or y) appears without a subscript it represents a set of vertices. When it appears with a subscript, e. g. , x_0, xl, x_2, etc. , then these are specific vertices where each xi refers to one and only one vertex. When referring to the vertex set, that is, the entire set of vertices in the universe of discourse, we shall use a capital X, e. g. , G = (X, T) or $X \underset{\Delta}{=} \hat{x}$ (x is a scientific paper).

In graph theory the upper case gamma, I' , is very popular. To facilitate the composition of this work the upper case letter T has been substituted for the gamma.

Explanation of Terms

The term paper, as opposed to the term document, will be used throughout. It more clearly illustrates the development of the theme of this research. It would, however, lead to no contradiction if the word document were substituted for the word paper.

The term citation has an ambiguity which needs clarification. The history of science, as all histories, builds on what has come before, and to that extent we can say that science is an edifice built from units or blocks of knowledge which we can call scientific papers. When a block of knowledge is added to the structure, we indicate which existing blocks are used to support the new addition, by providing a citation. When knowledge becomes "common knowledge" we no longer supply the pertinent citations. Each idea within a paper is dependent on some former idea, and to that degree should justly recognize the former idea by supplying a citation. However, what constitutes

[1] Table 5 in the Appendix contains a chart of the symbols used in this paper.

"common knowledge" varies with time, place and person. We are, therefore, forced to distinguish two types of citations, those which are implied, called implicit citations, and those citations which are clearly denoted, which we call explicit citations.

While it is important to realize that implicit citations exist, and it would be interesting to speculate on methods for converting implicit citations to explicit citations, this paper discusses only explicit citations. For brevity, the term citation will be used to stand for explicit citation.

CHAPTER III

CITATION INDEX TREATED AS A LARGE NETWORK

In order to treat a citation index as a graph, we must have: (1) a set X and (2) a function T mapping X into X. Here the set X is defined:

$$X \underset{\Delta}{=} \hat{x} \text{ (x is a scientific paper).}$$

The function, T, we shall call the <u>citing function</u>.[1] The expression Tx will mean "the set of papers which have cited the set of papers x." In Fig. 3.1, the paper x_0 is represented by a small circle, as are papers a, b, and c. The citing relationship is indicated by a directed line segment, sometimes called an edge, connecting the paper that is cited x_0 and the papers that do the citing, paper a, paper b, and paper c. An arrowhead is put on the line segment to indicate the direction of the action. The length of the line has no significance. Tx = (a, b, c}, reads "the set of papers represented by x is cited by the set of papers a, b, and c." In this example the set x, has only one member, namely x_0. This means that if one were to look at the bibliography of paper a, one would find that one of the entries would be paper xo. This is also true for the bibliographies or footnotes of paper b and paper c.

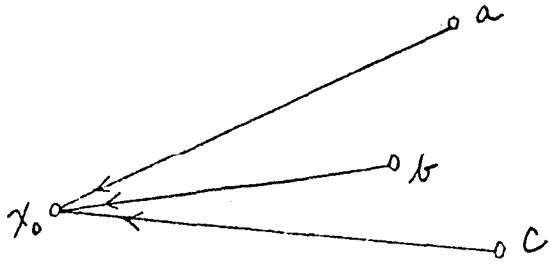


Fig. 3. 1 -- Example of Citing Function Tx_0

The expression, Tx, therefore provides a concise, non-ambiguous way of saying "the set of all papers that cite the set of papers represented by x."

The usual way of dealing with this notation is to provide another symbol, say y, which then can stand for the resultant set. Thus y = Tx. The symbol y, therefore, is an even more concise way of representing a set. For example, $y = Tx_0 = \{ a, b, c \}$.

The entire graph is denoted by G = (X, T). This states that the graph is composed of papers, and a function which relates these papers.

This notation permits us to make very precise statements which previously required considerable verbiage. In Fig. 3.2, we treat $x = \{ a, b, c \}$ so that the set y = Tx is represented by $\{ d, e, f \}$. Of course, x is Tx_0. Therefore $y = T(Tx_0)$. We define this as T^2x_0.

[1] Table 4 in the Appendix contains a representative sample of citing function expressions used throughout this work. Reference to the Table will supply the English language equivalent of the commonly used mathematical expressions.

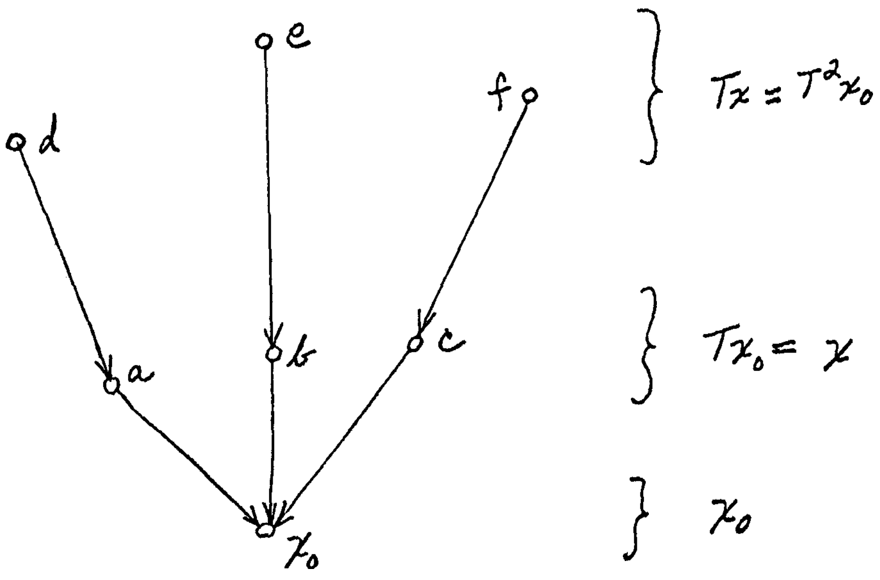


Fig. 3.2 -- The Citing Function as a Composite Function

By extending the use of the notation we get simplified expression such as T^3x, T^4x, T^5x . . to represent otherwise lengthy statements. For example T^3x represents "the papers that cited, the papers that cited, the papers that cited the set of papers x." Fig. 3.3 is an illustration of T^5x_0.

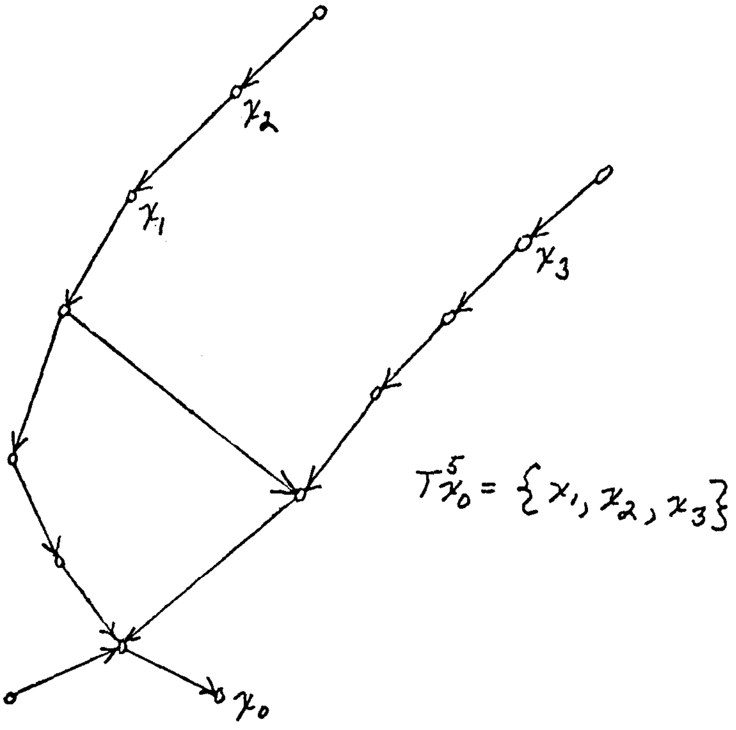
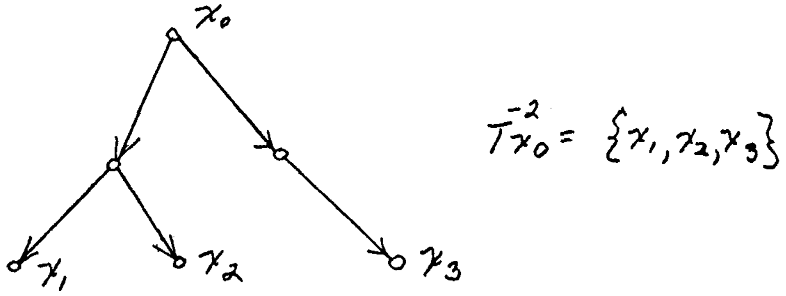


Fig. 3. 3 -- Example of T^5x_0.

By using negative exponents, for example, $y = T^{-1}x_o$, the inverse is equally facile to express. In this case, y is the set of all papers cited by x, that is, the bibliography of paper x_o. Fig. 3.4 is an illustration of $y = T^{-2}x_o$, that is, the bibliographies of the bibliographies of x_o.



Fig, 3.4 -- Example of the Inverse Citing Function

Types of Graphs

There are three fundamental types of subdivisions of a graph. They are called (1) subgraph, (2) partial graph and (3) partial subgraph. A subgraph of the graph G = (X, T) is defined as a graph of the form (A, Ta) where $A \epsilon X$ and T is restricted to A. For example,

$$Tx_o \wedge A = Tx_o \cap A,$$

This expression implies that we would like to consider all the papers which cite paper x_o; however, we would like to limit our discussion to only those papers which also belong to the set A. The requirement for membership in set A is that all the members shall have a certain property, for example, the paper must be published in a given journal or written in a certain language or the author must be a chemist or physicist, etc. We can write a mathematical description of all the papers which cite the set of papers x, including only those papers which have appeared in Journal of the American Medical Association. If we define

$$A \underset{\Delta}{=} 8 \text{ (A is a paper in the JAMA)},$$

Then the statement becomes

$$Tx \cap A.$$

A very practical expression which one would like to make in an information system might be: find all the references that would be of interest to me, but limit the output, at least initially, to only those journals which I have in my library. Another example would be to find all references to a set of papers but limit the output to papers published say between 1945-1965. Thus, we can generate a portion of a graph, Tx A, which is called a subgraph. It is important to note that all the edges and their directions are maintained in the resultant subgraph.

The second type of graph to be considered is the partial graph, sometimes called a section graph. In this case we impose a restriction on the citing function T. Thus, for the graph G = (X, T) let us consider (X, T_1) where $T_1x \subseteq Tx$ (for $T_1x \neq \emptyset$.) An example of a partial graph follows. Let us say that for some reason we would like to take the graph, G = (X, T) and generate the graph containing the vertex set, X of G but limit the citing function to only those edges which go from a to b, where a and b are of the same type. Formally, the graph G will now be of the form $G' = (X, T_1)$ where $T_1x \subseteq Tx$ for all x. G' is the partial graph that results if we limit our discussions only to edges which connect vertices that have a given property, that is, the journals are the same, or some similar property.

The third type of graph is the <u>partial subgraph.</u> In this type of graph we impose a restriction on the vertex set X, and we also impose a restriction on the citing function, T. In other words, the vertices, X, have limits; and the citing function, T, is limited also by the type of citing. The graph, G = (X, T) with the following restrictions:

(1) $A \subseteq X$

(2) $T_1x \subseteq Tx$ A

will result in the graph, $G' = (A, T_1)$.

This would be useful in the following example:

Let G = (X, T) where

$X \underset{\Delta}{=} \hat{x}$ (x is a scientific paper)

and T is the citing function mapping X onto X.

Let X be restricted to the class A, and let the citing function T be restricted to the T_1-type of citation.

$A \underset{\Delta}{=} \hat{A}$ (A is a paper by an author who is a chemist)

and T_1 is a citation where the journal of the paper doing the citing, and the journal of the paper getting cited, are the same, let us say <u>Journal of Organic Chemistry.</u>

$A \underset{\Delta}{=} \hat{A}$ (A is a paper by an author who is a chemist)

$B \underset{\Delta}{=} \hat{B}$ (B is a paper which has appeared in <u>Journal of Organic Chemistry</u>)

$T_1 \underset{\Delta}{=} \hat{x}\hat{y}\ (y = Tx\ \&\ y \epsilon B\ \&\ x \epsilon B)$ [2]

Thus we would like a citation network generated with the following restrictions:

(i) $X \wedge A$

(ii) $Tx \wedge T_1x$

This is an example of a partial subgraph.

Thus it becomes clear that there is a myriad of citation networks possible for analysis. All one has to do is let the restriction A in X A change and we can generate a very large number of citation networks. Moreover, one can hold X constant and alter

[2] When we say that a restriction has been placed upon the edges of a graph we mean that the restriction is placed pairwise on the two vertices the edge connects. This equation defines a citing function, T_1, as a set of edges which is a set of ordered pairs, $\hat{x}\hat{y}$, where the x and y have the indicated characteristics.

the meaning of T, the citing function, $(T_1, T_2, T_3 \ldots Tn)$ and this too can generate a very large number of citation networks. The first is the number of subgraphs possible, and the second is the number of partial graphs possible. Their product is the number of partial subgraphs possible. It becomes one of the major objectives of network analysis to obtain the partial subgraph that is the most meaningful to the user. This is facilitated only when the user can precisely specify his needs.

Formation of Citations

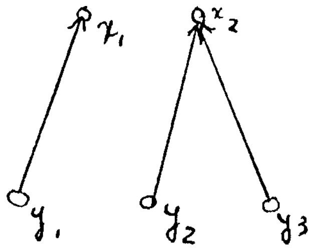
When an author provides a citation to indicate his use of the literature, the citation is in essence a description of the paper which he is attempting to uniquely identify. Because there is no universally accepted method for the formation of citations, the description which the author provides is to a great degree a matter of individual taste, and as such is subject to the whims of the author's personality. In many cases this results *in* vague and ambiguous citations not to mention the citations which are overtly incomplete and incorrect. This is clearly seen by the examination of a number of bibliographies of papers. As a result of entering such citations into an information system, the system may very well contain many different descriptions of the same paper. Consequently, any given paper may be represented in the network by more than one vertex.

Another aspect of processing citation data which may add "extra vertices" to the network, is when a citation is to a specific page of a paper, rather than to the paper as a whole,, where the whole paper is identified by its page range. In actuality, the citation may very well be to a sentence in one of many paragraphs on the cited page. If one considers the page number when constructing a citation network a paper may be represented by more than one vertex. Moreover, the number of different possible vertices representing a paper will be the number of different pages on which the author is cited, This may be an advantage or a disadvantage depending on the situation.

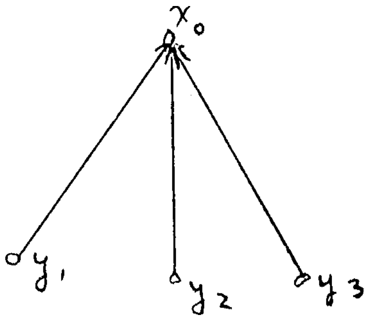
The two main causes of extra vertices in the graph are thus: (1) inexact use of citations by authors, and (2) use of specific page rather than the page range. The recognition of the existence of these extra vertices permits us to distinguish two types of graphs -- the first is a network of bibliographic entries, and the second is a network of papers. In passing, we mention that correcting citations is possible and has been done by mechanized procedures. The problem which the page range presents will be discussed more thoroughly below.

Bibliographic Network vs. -Network of Papers

A citation index network is a network of bibliographic entries. When a bibliographic entry is entered into a citation index system as a source item, the entry will contain the page range or some appropriate designation for a preferred page identification. However, when the paper is entered into the system as a reference item, it is difficult, if not impossible usually, to enter the page range or the preferred pagination, especially if it is not known. The author when providing citations usually provides a specific page number, usually the first page, rather than the page range of the journal article. In Fig. 3. 5a, xl and x2 are bibliographic entries which have been entered into the system as reference items. However, x_1 and x2 are merely variations of the source entry, x_0, the preferred bibliographic entry for the paper. Fig. 3. 5b illustrates the graph which results from combining vertices which belong to the same papers. Only when we have one and only one entry in a network for each paper, can we then say that we have a network of papers. Let us state this in another manner. If there exists n variations in the citing of paper x_0, call these variations x_1, x2, x3, . . . xn. Each of these variations of x_0 have resulted from the entry appearing in n or more bibliographies. When the entry for the source document, x_0, is entered into the system, we



Network of Bibliographic Entries
Fig. 3. 5 a



Network of Papers
Fig. 3. 5b

may now have as many as n+l representations of the same paper in the system. It is assumed that the entry created from the source document is the preferred entry. It therefore remains to identify all the other n forms of x_0, and convert them to the preferred form, x_0. In general, we must form subsets where the members of each subset are different forms of the same bibliographic entry. Moreover, these subsets form a partition, so that after the partitioning process only one entry remains in the system for each subset. All the edges between vertices are maintained. The resultant subsets are mutually exclusive and jointly exhaustive.

To summarize, the term bibliographic entry or entry for the purpose of this work has a many-to-one relationship to the term scientific paper or more simply paper. This is plausible if we recall that a paper has a page range, while a bibliographic entry could be more specific.

Elimination of the Ambiguity from the Nodes of the Citing Function

The statement $y \varepsilon Tx_0$ is ambiguous. Graphically it could mean Fig. 3. 6a or Fig. 3. 6b.

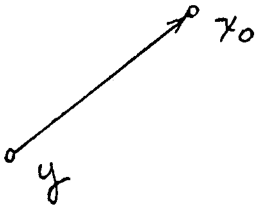


Fig. 3. 6a -- Sample Graph

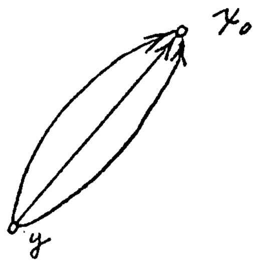


Fig. 3. 6b -- Multi-Edged Graph

Usually this distinction is disregarded. In this analysis, however, such imprecision would lead to confusion, especially when discussing bibliographic coupling. We shall adopt the convention, therefore, that ordinarily the variables x and y will stand for a paper unless otherwise stated. Under these circumstances $y \varepsilon$ Txo will be denoted by Fig. 3. 6a. It is necessary to be consistent in our assignment of the designation of vertices. Fig. 3. 6c, for example, denotes that papers y_1, y_2, y_3 all have cited x_0. However, in the case where $y_1 = y_2 = y_3$, that is to say, in the case where y_1, y_2, and y_3, are the same source document we would prefer that the source document y_0, be denoted by one vertex rather than three. This would result in Fig. 3. 6b. The diagram in Fig. 3. 6c can be replaced by the diagram in Fig. 3. 6d. For this reason with each edge (y_0, x_0), i. e. , the path going from y_0 to x_0 there is associated a number $n(y_0, x_0)$

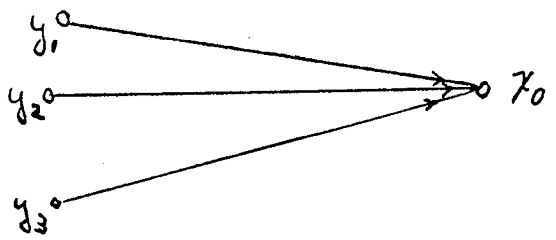


Fig. 3. 6c -- Sample Graph

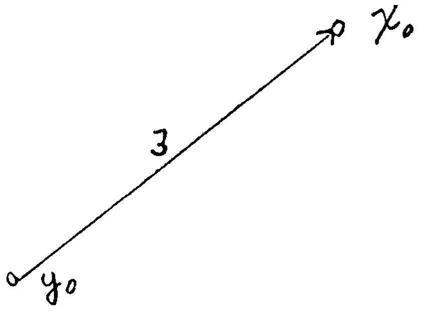


Fig. 3. 6d -- Multi-Edged Graph

which is the number of times a paper yo has cited paper xo. Thus we have n = n(yo, x_0). In Fig. 3. 6d, therefore, we have n = $n(y_0, x_0)$ = 3. We use a number to indicate the number of edges rather than drawing them. The importance of this distinction will become apparent when we discuss transformation graphs in Chapter VI (p. 32).

In order to avoid confusion it is necessary to indicate at the outset the following special case. When an author cites his own work it is called a self-citation. A special case of self-citation is when the entry in the bibliography is to the paper in which it is appearing. How is this special case represented in a citation network? Ordinarily, we have a directed line segment going from the paper that is doing the citing to the paper that gets cited. In this case, however, both these papers are the same. Consequently, the directed line segment is drawn from the citing node back to itself. This is called a loop. Fig. 3.7 paper x_0 cites itself; and paper x_1 cites itself and paper xo.

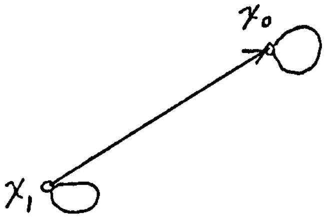


Fig. 3.7 -- Graph with Loop

In this context, including the arrowhead on the loop is arbitrary since it adds nothing to the representation of the self-citation.

Checking the Citation Network for Validity

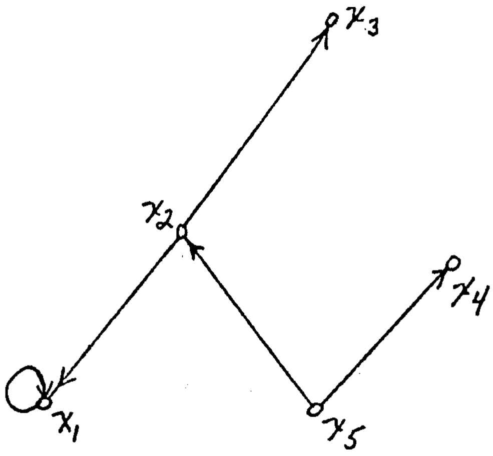
Very often reference is made to papers which have not as yet been published. Illustrative of these types of citations are: "in press," "to be published," "unpublished manuscript," etc. For these articles a date of publication is usually not provided. For the purpose of the discussion in this paper, these citations will be disregarded. This permits us to develop a method which can easily verify the construction of a citation network. If a directed line is drawn from a to b we can say that the date of paper a is more recent than the date of paper b. An examination of each pair of papers in the network constitutes a simple procedure for validating the chronological organization of the network. Although this is not a final test, it is a simple procedure for eliminating gross errors.

In summary, one can imagine the set of all the world's scientific papers as one enormously large directed graph of the form G = (X, T). Any graph that is less than this is a subgraph, a partial graph or a partial subgraph. It is imperative to have a clear understanding of not only what is present but also of what is absent from the network. More correctly, it is important to know what type of graph is presented and consequently, what are the restrictions and limitations that have been placed on the original graph from which the present graph was generated. These comments apply not only when the citation index is in a graph form but also when it is in the form of a columnar index, as it is in the Science Citation Index. In passing, we may note that one does not necessarily have to have all the world's scientific papers in a citation index network, for it to be of value. To have a graph that is something less than the total graph may not only be useful but may be preferred.

A Matrix Can be Used to Represent a Graph

The use of matrix theory as a facile means of representing graphs, especially large graphs, is possible through an analysis of the elementary notations of matrix theory.

A matrix is an array of terms, composed of columns and rows. Initially we will only be interested in matrices where the number of columns and rows are the same. This is called a square matrix. A term in the matrix can be indicated by a^1_j, where i is the row of the term and j is the column of the term. For example, in Fig. 3. 8, $a^5_2 = 1$.



$$A = \begin{array}{c|ccccc|c}
(j = & 1 & 2 & 3 & 4 & 5) & \\
 & 1 & 0 & 0 & 0 & 0 & (i = 1 \\
 & 1 & 0 & 1 & 0 & 0 & 2 \\
 & 0 & 0 & 0 & 0 & 0 & 3 \\
 & 0 & 0 & 0 & 0 & 0 & 4 \\
 & 0 & 1 & 0 & 1 & 0 & 5)
\end{array}$$

Fig. 3. 8 Matrix Representation of a Graph

For our purpose we let the rows represent the set of papers in the graph, and we also let the columns represent the papers in the graph. If $|X| = 5$ as it does in the above example, then the total number of cells or terms in the matrix is 25. Each cell of the matrix contains a symbol which indicates the relationship between the row and column elements. In our context i is a paper that does the citing, and j is the paper that is cited. . If paper xi cites paper xj we have $x^1_j = 1$. If paper xi does not cite paper xj , then $x^1_j = 0$. These are the only two conditions which can exist. $_5$In Fig. 3.8 paper x2 cites x3 so that in the matrix $x^2_3 = 1$; paper x5 cites x4, so that x4 = 1. Paper x3 does not cite x_1, so that $x^3_1 = 0$. It is interesting to note that paper x_1 cites itself, so that $x^1_1 = 1$. In general, the x^1_j terms, where i = j form the principle diagonal. Inspection of the principle diagonal for the terms which have a 1, indicates which vertices have loops. In Fig. 3. 8 there is only one term on the principal diagonal which is equal to 1, x^1_1; this term is the only vertex in the network which has a loop.

The size of the matrix depends on the number of vertices in the graph. If we impose restrictions on the set of vertices in a graph, as we do in a subgraph, the associated matrix will be smaller and in general easier to handle then the original graph. However, working with a partial graph does not affect the size of the matrix. Representation of partial subgraph will also affect the size of the matrix.

As the discussion proceeds the value of the matrix representation will become clear. To summarize, a matrix and a graph can completely represent a citation index network. The use of a graph or matrix depends on the circumstances. The terminology used depends on the type of representation discussed. Certain terms, in a given context, are completely interchangeable. Thus, if we have a citation index represented by a graph, and the graph represented by a matrix, the relationship of the terms are illustrated in TABLE 1.

TABLE 1

THREE REPRESENTATIONS OF A CITATION INDEX

Form	Tabular	Graph	Matrix
item	citation entry	vertex	term a_j^i
the relation "cited by"	one listed under the other	a directed line from one vertex to another	1 in cell a_j^i
the relation "not cited by"	no entry	no line	0 in cell a_j^i

Matrices for Citation Network Representation

It is presently commonplace for computers to manipulate matrices quickly and efficiently. Many of the large-scale computers are binary machines. They are, therefore, especially useful for representing matrices efficiently since each bit in the machine can be related to a cell in the association matrix. In the type of matrices discussed above each cell contains either a one or a zero. This will correspond to a bit in the machine being either "on" or "off."

The ability to represent a citation network as a matrix, therefore, permits us to make use of existing programs and procedures for analyzing citation index data in matrix form.

Programming Language for Citation Index Structures

All the material presented in this paper is independent of any machine programming language. The statements in graph theory notation can be converted to almost any programming language with varying degrees of difficulty. We note that the programming language of Iverson[3] appears applicable to graph structures. Sussenguth [4] uses Iverson's language to process graphs in the form of trees.

[3] K. E. Iverson, A Programming Language, John Wiley and Sons, Inc., New York (1962).

[4] E. H. Sussenguth, Jr., "Structure Matching In Information Processing," Ph. D. dissertation, Dept. of Applied Mathematics, Harvard University (1964).

CHAPTER IV

SEARCHING THE CITATION INDEX NETWORK

The analysis presented in this section assumes that a machine-form citation index is available. Since there do indeed exist files of millions of citations in machine-form, it is the next logical step to search these machine-form files making use of computer techniques. The following discussion shall indicate that there will be two fundamental types of information retrieval systems with regard to machine-form citation index structures.

In most information retrieval systems the user is at a distance from the system. The user makes his needs known to the system through a person who transforms his question into a statement which is non-ambiguous, and which is designed to be compatible with the logic and organization of the programs and computer of the information retrieval system. This ordinarily means that the intermediary person, let us call him the coder, has a knowledge of the user's needs and a knowledge of the information retrieval system. The coder enters the query into the system. While queries are usually batched, this is not of primary importance. Once the query is entered into the system, the files are searched for records within the system which satisfy the query. Usually this procedure will be a type of matching operation. In an information retrieval computer system the fundamental operation which is executed within the computer is the compare operation. If there is a match or a number of matchings, they are accumulated, formated and produced as output of the search -- the search product. During this entire process the content of the query has not changed. There has been no communication between the system and the original user. There has been probably no contact with the coder. Through the entire process the search query has remained constant. Most of the systems in operation today are of this kind. While there are good reasons why this is true, most of which are economic, it is not obvious that this should necessarily be the case. Some of the more sophisticated systems have built into the query, by the coder, a hierarchy of "relaxations. " Very briefly, the query might demand from the system, items matching five descriptors. If the system is interrogated and it is found that there are no records which satisfy this condition, it would indeed be an inefficient system which completed the search at this point. The more efficient systems permit the user, usually the coder, to specify which of the descriptors are least important so that in the event nothing is found which satisfies all five demands, the program knows to look next for records which have four of the five descriptors. This line of reasoning can be extended so that the demand on the user, the coder, and the system increases to the point where it becomes too cumbersome to be feasible. An example of this type of system would be the MEDLARS system wherein if we extended the demands of the search query formation on the user and coder, the system would become more and more taxed. It is important to point out that in the MEDLARS system example, the queries are batched so that all the necessary demands on the user and the coder must be satisfied in advance of the running of the programs against the query through the computer.

The next step in the overall approach of system design, is to make the search product more compatible with user's needs by bringing the user (or the coder initially) in intimate contact with the system. This can occur only under the following conditions, The user will need a device to facilitate communication with the computer system. This can take the form of a display-tube console equipped with a light pen. The display is used for output and the light pen is used for input. Another available device is a typewriter-keyboard console. The latter is more practical, less expensive, and at present, has received greater acceptance. In addition to the system's ability to communicate with these remote devices, the computer must have the capacity to perform simultaneous processing of programs. It must be a time-shared computer system. The MAC system at MIT is a time-shared system, where the remote typewriter consoles are scattered about the MIT campus. This large-scale system makes use of random access devices which can store vast quantities of data and programs. These types of systems have been sufficiently documented, so that further discussion would be repetitious.

The type of system described above is usually called a man-computer system. Many authors have discussed this type of system in depth. Riesner[1], however, has used this type of system in a unique manner. Her research includes the use of a man-computer system for the storage and on-line interrogation of thesauri. Briefly, her work describes the procedure of storing a thesaurus in a computer and then using a remote console for interrogating the thesaurus for such needs as: look-ups, updating, checking for consistency, etc. She has called this type of thesaurus the "man-computer thesaurus. " In similar fashion, we can extend this line of reasoning to include a "man-computer citation index. " In short, using an on-line console for display of citation index data is now possible. The most important aspect of the system is that the user (or coder) has the ability to change his query based on the output he receives. The key concept is that the system provides feedback. The user is now in the system; he is a part of a loop of the system. This should be contrasted with the systems which required batching, described earlier where the user was not part of the system. These two types of systems of search, the conventional batching system, and the man-computer system, have different needs as far as the manipulating of citation index files are concerned. In each system there is a different set of variables. In the batching system all decisions must be made in advance, while in the man-computer system, since there is feedback, many decisions can be made as the search progresses.

The design of a "man-computer citation index system," or merely an efficient batching search system, depends on the development of a facile notation which has the ability to specify the search query succinctly, clearly, and unambiguously. The following presentation, which is made possible by the use of graph theory, will attempt to analyze the techniques which have been discussed in the literature which could be used for search query formulation. Both the methodology and terminology are rigorously defined.

Bibliographic Coupling

The purpose of this section is to apply the graph theory notation developed in the earlier sections, to the concept of bibliographic coupling. If we can find a mathematical description of bibliographic coupling we are then in the advantageous position of being able to compare other techniques to it, and thereby realize the differences and similarities. This will of course result in a rigorous definition of bibliographic coupling.

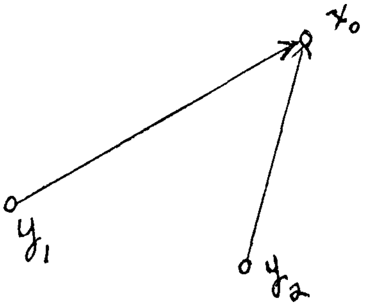
Below is cited an example of bibliographic coupling by Kessler:

> "In this study, it is postulated that a number of scientific papers bear a meaningful relation to each other (they are coupled) when they have one or more references in common. We define a unit of coupling: Two papers that share one reference contain one unit of coupling. A

[1] P. Reisner, "A Note on Minimizing Search and Storage in a Thesaurus Network by Structural Reorganization of the Net, " Final Report on Contract AF 19 (628)-2752, pp. 365-373 (1964).

> coupling criterion is defined by the combination of coupling units between two or more papers. We define two criteria of coupling:
>
> "CRITERION A: A number of papers constitute a related group GA if each member of the group has at least one reference (one coupling unit) in common with a given test paper P_0. The coupling strength between P_0 and any member of GA is measured by the number of coupling units between them. G_A^n is that portion GA that is linked to Po through n coupling units. (According to this criterion there need not be any coupling between the members of GA, only between them and Po.)
>
> "CRITERION B: A number of papers constitute a related group GB if each member of the group has at least one coupling unit with every other member of the group. The coupling strength of GB is measured by the number of coupling units between its members. Criterion B differs from Criterion A in that it forms a closed structure of interrelated papers, whereas Criterion A forms an open structure of papers related to a test paper. "[2,3]

Two papers have one unit of coupling if they share one reference in common. An example of this is illustrated in Fig. 4.1.



WHERE $y_1 , y_2 \in T_{x_0}$

Fig. 4. 1 -- Two Papers With One Unit of Coupling

From this we can say that if $|T^{-1}y_2 \cap T^{-1}y_1| = 1$, then y_1 and y_2 contain one unit of coupling strength. In other words, the number of members of the intersection of the bibliographies of papers y_1 and y_2 is equal to 1. If this number was n, papers y_1 and y_2 would have n coupling units.

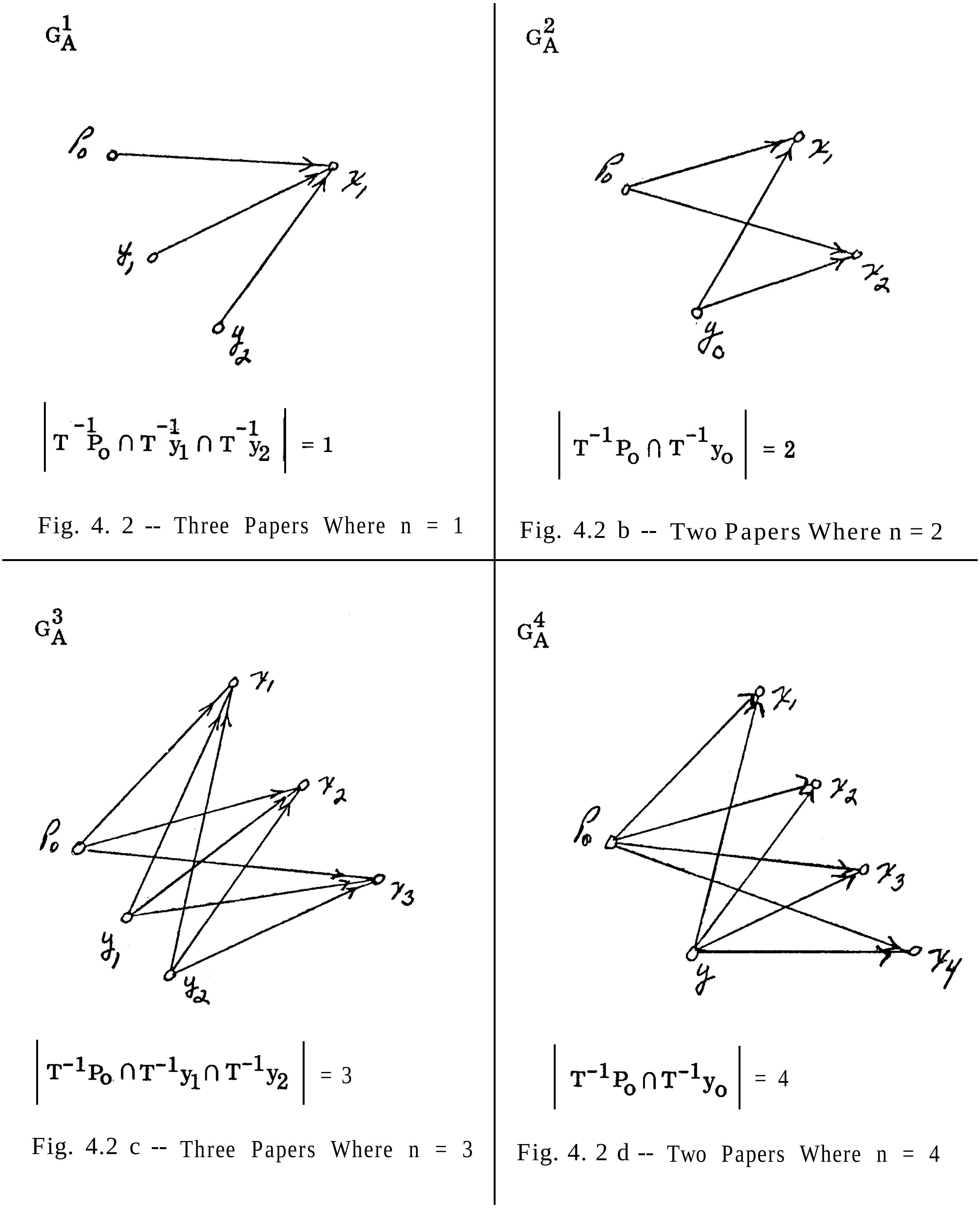
We now describe two criteria of coupling. Criterion A is established if we let $y_2 = P_0$ where P_0 is a test paper, and generate all the sets for which the right hand side of the above equation has the value 1, 2, 3, ... Therefore, G_A^n is the set of papers where $|T^{-1}P_0 \cap T^{-1}yi| = n$. This creates a partitioning of the graph G, based on the value of the number n.' It is clear, therefore, that $\bigcup_{k=1}^{n} G_A^k = GA$ while $\bigcap_{k=1}^{n} G_A^k = \emptyset$ (where P_0 is not included.)

[2] M. M. Kessler, "An Experimental Study of Bibliographic Coupling Between Technical Papers, " IEEE Trans. Information Theory, IT9:49 (1963).

[3] In an attempt to be consistent with the quotation from the work of Kessler, I have used his symbolism in this chapter to explain his ideas. GA and GB, therefore, are not graphs as they are in all other chapters of this work, but are groups of references. This is the only chapter in which this notation will be used. When P_0 is used this implies that we are using Kessler's symbolism.

In order to increase the visualization of GA, Fig. 4.2a-d illustrate G_A^n for n equal to 1, 2, 3, and 4.

Fig. 4.2a-d -- Examples of Bibliographic Coupling



$$\left| T^{-1}P_o \cap T^{-1}y_1 \cap T^{-1}y_2 \right| = 1$$

Fig. 4. 2 -- Three Papers Where n = 1

$$\left| T^{-1}P_o \cap T^{-1}y_o \right| = 2$$

Fig. 4.2 b -- Two Papers Where n = 2

$$\left| T^{-1}P_o \cap T^{-1}y_1 \cap T^{-1}y_2 \right| = 3$$

Fig. 4.2 c -- Three Papers Where n = 3

$$\left| T^{-1}P_o \cap T^{-1}y_o \right| = 4$$

Fig. 4. 2 d -- Two Papers Where n = 4

We can also define G_A^n as $\left|T^{-1}P_o \cap T^{-1}y_o\right| \geq n$ so that G_A^n includes all G_A^m where $m > n$. For Example, $G_A^5 = G_A^5 \cup G_A^6 \cup G_A^7 \ldots$ or $G_A^5 = \bigcup_{k=5}^{n} G_A^k$. In this sense of the definition we do not have the partitioning which previously occurred. It is in this latter sense that the concept of G_A^n is usually applied. Fig. 4. 3 is an illustration of G_A^2 in this latter sense.

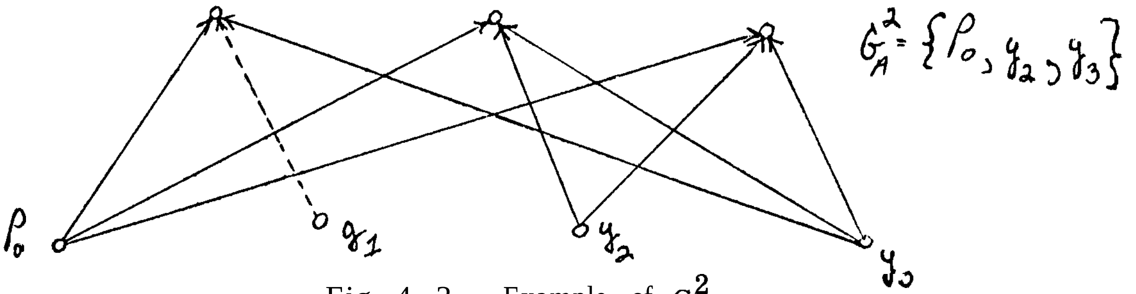


Fig. 4. 3 -- Example of G_A^2

The dotted lines indicate that $\left|T^{-1}P_o \cap T^{-1}y_1\right| = 1$. In this illustration paper y_1, has been omitted from G_A^2 because it did not have the sufficient number of references in common with P_o.

Criterion B is nothing more than $\left|T^{-1}y_i \cap T^{-1}y_j\right| \geq n$ for all i and j. The significant difference here is that every paper is linked with every other paper.

Fig. 4.4 is an illustration of a case of G_B^2 where $\left|T^{-1}y_1 \cap T^{-1}y_2\right| = 2$ and $\left|T^{-1}y_2 \cap T^{-1}y_3\right| = 2$ and $\left|T^{-1}y_1 \cap T^{-1}y_3\right| = 2$. $G_B^2 = \{y_1, y_2, y_3\}$.

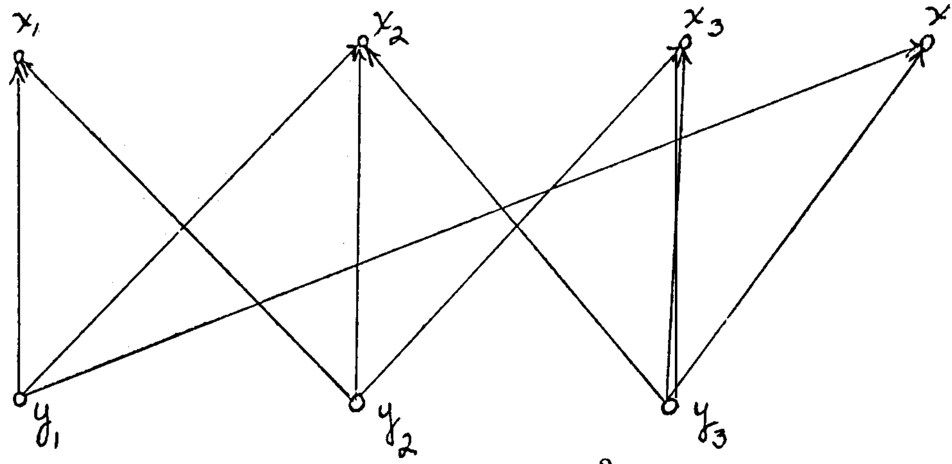


Fig. 4.4 -- Example of G_B^2

Criterion B is much more demanding than Criterion A, and consequently it is easy to believe that GB papers are more closely related than are GA papers.

Two papers bear a meaningful relationship to each other if they are bibliographically coupled. However, the graph theoretic notation illustrates, in the following paragraphs, two aspects of bibliographic coupling which suggest that further research is necessary.

Given $G_A^2 = \{y_1, y_2\}$ where $|T^{-1}y_1| = 2$ it would appear that there is a significant difference if $|T^{-1}y_2| = 2$ or $|T^{-1}y_2| = 25$. To say that these two papers have the same "coupling strength" leaves much to be desired.

Furthermore,4 no comment has been made in the literature about the following situation. Let $G_A^* = \{P_o, y_o\}$ and consider the situation where $T^{-1}P_{o1}$ contains many entries where say k of them are the same. If k = 4 we can have $T^{-1}P_o = \{xl, x2, x3, x_4, \ldots\}$ and $x_1 = x2 = x3 = x4$. Under these circumstances, if the coupling papers are x_1, x2, x3, x4, it would seem that there is a "stronger coupling" than if $x_1 \neq x2 \neq x3 \neq x4$ (See earlier section where distinction is made between paper and bibliographic entry.) If the strength of relationships between papers are to be related to a number of entries, there must be a statement about the condition when the entries are identical.

Cycling

The purpose of this section is to apply the graph theory notation developed earlier to the terminology Garfield used in his discussion of (i) search strategies and (ii) citation indexes in sociological and historical research. In this section we discuss only the definitions of terms that can be applied to search strategies, and while these same terms can be useful for sociological and historical interpretation, we shall reserve this material for a later section.

Garfield has discussed search strategy for citation indexes as follows:

> "It is significant that, for similar reasons, when using citation indexes for literature searching, the search strategy frequently requires examination of references not only to a specific paper. It may frequently be necessary, if not preferable, to examine references to the papers in the bibliography of the target paper. This technique is called cycling. If S are the sources which cite one or more members of the bibliography B in target reference R, S(B(R)) is the pertinent list of papers. "[4]

Converting these comments into citing function terminology we let R = x, the target reference. B(R) therefore becomes $\overset{-1}{Tx}$, the bibliography of the target reference. If S are the sources that cite the member of B(R) these now can be written $\overset{-1}{T}(Tx)$. In general, we define cycling as the successive application of the citing function. -der to keep the definition general we need not specify that the target reference be a single paper; it may very well be a set of papers. Table 2 contains examples of cycling written in English and in citing function notation.

It should be noted as a technical consideration that when the exponents of the successive application of the citing function are the same, they are additive, as in example (i) and (iv). Otherwise they are not additive as in (iii). Ordinarily, $\overset{-1}{T}(\overset{-1}{T}(Tx)) \neq \overset{-1}{T}X$ for

$$\overset{-1}{Tx} \subseteq \overset{-1}{Tx}\ \overset{-1}{(T(Tx))}$$

[4] E. Garfield, "Citation Indexes in Sociological and Historical Research," p. 290.

TABLE 2

EXAMPLES OF CYCLING IN ENGLISH AND IN CITATION FUNCTION NOTATION

English	Citing Function Notation	
The set of papers that cite the papers that cite x.	$T^2x = T(Tx)$	(i)
The set of papers that cite the entries in the bibliography of xo.	$T(T^{-1}x_0)$	(ii)
The papers in the bibliographies of the papers that cite the papers that cite the papers in the bibliographies of x.	$T^{-1}\left(T\,(T^{-1}x)\right)$	(iii)
The papers in the bibliographies of the papers in the bibliographies of x.	$T^{-2}x = T^{-1}(T^{-1}x)$	(iv)

The order of application of the citing function is ordinarily important.

$$T^{-8}(T^{2}(T^{-1}x)) \neq T^{2}(T^{-8}(T^{-1}x))$$

From Fig. 4. 5 we see that

$$T^{-1}x_1 = \{y_1, y_2\}$$

and $T(T^{-1}x_1) = \{x_1, x2>$

and $T^{-1}(T(T^{-1}x_1)) = \{y_1, y_2, y_3\}$

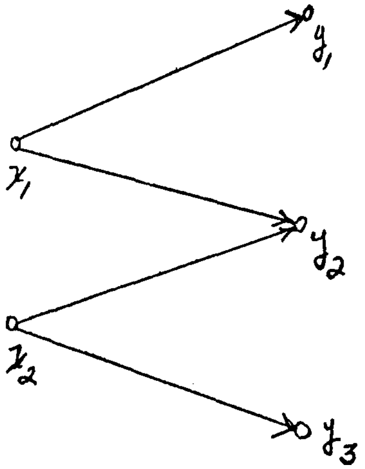


Fig. 4. 5 -- Exponent of the Citing Function not Always Additive

Clearly then $T^{-1}x_1 \neq T^{-1}(T(T^{-1}x_1))$ for y_3 is not included in the left hand side of the above equation.

We can further extend the concept of cycling by imposing restrictions on both the citing function and the sets being operated upon. At each set generation we can, for example, eliminate those members which do not satisfy a given criterion. Similarly, with the application of the citing function, we can modify its meaning in n ways giving $T_1, T_2, T_3 \ldots T_n$; this would of course also limit the number and kind of members of the sets generated during the cycling process. Below is an example of the cycling process with the given limitations. Let T_1 be references from journal A to journal A (written in language C), and let T2 be references from language B to language C and let the set generated be limited to papers written after 1960. Therefore, $T_2(T_1^{-1}x)$, is a search we might want to perform if we were interested in the recent use of papers in journal A which were frequently used by journal A, and we were interested in their use by journals in language B. An advantage of using this notation is that the negative exponent takes you back in time, while the positive exponent takes you forward in time, a very natural way to distinguish the results of the citing function. If we define x as the set of papers which were written by geneticists or written in genetics journals, we can define the Genetics Citation Index as Tx. In order to indicate that the original papers x, appear in the index we would write $\{x\} \cup Tx$. Needless to say this is a description of the content of the Genetics Citation Index and not its organization.

Cycling and Coupling -- Similarities and Differences

Let us assume that we have a network as illustrated in Fig. 4. 6. From the point of view of bibliographic coupling we can say that

$$T^{-1}x_1 = \{y_1, y_2\}$$

$$\text{and } T^{-1}x_2 = \{y_2\} \text{ and } T^{-1}x_3 = \{y_2, y_3\}$$

therefore $\left|T^{-1}x_1 \cap T^{-1}x_2\right| = 1$,

$\left|T^{-1}x_1 \cap T^{-1}x_3\right| = 1$ where $P_c = x_1$ so that $G_A^2 = \{ x2, x3$

and $\left|T^{-1}x_1 \cap T^{-1}x_2 \cap T^{-1}x_3\right| = 1$ so that $G_3^1 = \{x_1, x2, x3\}$

From successive use of the citing function we can say that:

$$T^{-1}x_1 = \{y_1, y_2\}$$

$$T(T^{-1}x) = \{x_1, x2, x3\}$$

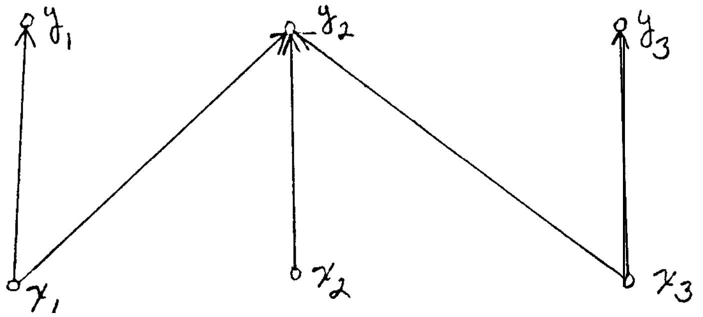


Fig. 4. 6 -- Sample Network

The difference between cycling and bibliographic coupling is clear when we realize the purpose behind each concept. In general, cycling is broader, more general than bibliographic coupling. With the cycling process one may start with one paper and generate all the papers which satisfy the cycling configuration of citing functions. In this sense, therefore, the cycling process is relatively unlimited in its power to retrieve entries. The major advantage of cycling is that it gives the user the ability to search through the network in any direction in time. While some of the fundamental strategies are clear, it remains to be shown which search strategies are the most fruitful. This can only be done through extensive experimental studies with an existing system.

Using coupling as a searching device, the problem is compounded since one must start with a test paper and then search through all the other papers in the file to find those papers to which the test paper is bibliographically coupled. The two methods appear very different. This is true in theory. But in practice, when working with a very large network after starting with a test paper, one would then locate all the other x's by finding Ty where y = Tx^{-1}. In the practical situation, therefore, the technique of cycling and bibliographic coupling are similar in their objectives. Both assume that a citation index network is available. In the coupling process one goes on to build a hierarchy of relatedness between x and the test paper in Criteria A and a similar type of hierarchy in Criteria B. We can conclude that the cycling process is especially equipped to retrieve from a network of entries a particular group of entries which satisfy a given criterion defined by the successive application of the citing function. When we start with a target paper or test paper the organization of these retrieved documents may be arranged, if desired, by the bibliographic coupling process. These comments are summarized in Table 3 (see page 28).

Combining the Citing Function with Boolean Notation in the Formation of Search Query

The use of Boolean notation for the formulation of the search query as it is done in the MEDLARS system can be expanded to accommodate the use of the citing function in the cycling process, and thereby specify search criteria concisely. A search for the papers which cite papers x_1 or x2 or x3, can be written $T(x_1 \cup x_2 \cup x_3)$. This example indicates that batch processing can be possible, for if necessary we can combine a number of search requests into a single statement. The computer programs would then optimize the method for locating all the appropriate bibliographic data. The only restriction on the number of successive applications of the citing function and on the number of Boolean operators would be the limitation of the object machine.

TABLE 3

COMPARISON OF CYCLING AND BIBLIOGRAPHIC COUPLING

Process	Start with	Goal	Technique
CYCLING	target paper(s)	retrieve papers meaningfully related to target papers	successive application of the citing function
BIBLIOGRAPHIC COUPLING			
Criterion A	test paper	retrieve papers meaningfully related to a test paper P_o	1) application of the citing function in order to avoid checking every entry in the file 2) arrange by strength of coupling.
Criterion B	single member of the aspired group	retrieve group which are mutually related	1) application of the citing function to locate related papers 2) arrange by strength of mutual coupling

CHAPTER V

LOCATING PATHS IN THE CITATION NETWORK

The matrix representation is very efficient for finding powers of the citing function within a group of papers. For example, let X = {xl, x_2, x_3, x_4} and let G = (X, T) be the graph in Fig. 5. 1.

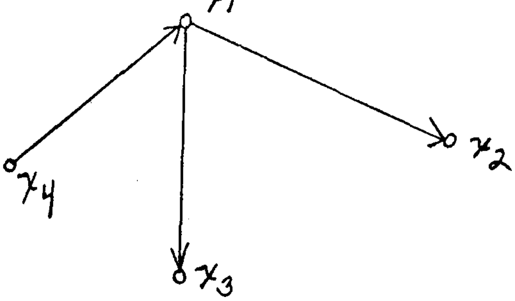


Fig. 5. 1 -- Sample Graph

Associated with this graph is the matrix A in Fig. 5.2

$$\begin{array}{c|cccc} & x_1 & x_2 & x_3 & x_4 \\ \hline x_1 & 0 & 1 & 1 & 0 \\ x_2 & 0 & 0 & 0 & 0 \\ x_3 & 0 & 0 & 0 & 0 \\ x_4 & 1 & 0 & 0 & 0 \end{array} = A$$

Fig. 5. 2 -- Association Matrix

In Fig. 5. 3 the square of the matrix is illustrated. The resultant matrix A^2., which is obtained by matrix multiplication, has $x_4^2 = 1$ and $x_4^3 = 1$ which means that $x_4 \varepsilon T^2 x_2$ and $x_4 \varepsilon T^2 x_3$. The A^2 matrix, therefore, clearly illustrates all papers which have the T^2 relationship. If we continue we see that $A^3 = 0$ so that $T^3 x = \emptyset$ for all $x \varepsilon X$. To

$$A^2 = \begin{vmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{vmatrix} \circ \begin{vmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{vmatrix} \qquad \begin{vmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{vmatrix}$$

Fig. 5. 3 -- Square of the Association Matrix generalize, we may say that to locate the longest path in the graph we locate $A^n = 0$ where $A^{n-1} \neq 0$ and the cells in A^{n-1} which have l's are the longest path(s) in the graph. In passing, we may note that if $|X| = 5$, the longest path cannot exceed 4 so that the (n-l) in the expression A^{n-1} will always be less than or equal to the number of elements in the graph.

As a technical consideration it will be noted that the above procedure will not work unless the cells on the principle diagonal are all zero. It is therefore necessary to remove the loops from the graph before converting to the matrix representation in this context.

Critical Path

In CPM[1] and PERT[2] the term critical path is of primary importance. When using these scheduling and planning techniques, the critical path is the path along which a delayed activity holds up the completion of the project. The critical path is the backbone of these network techniques. In these methods there are two vertices which are of primary importance -- the initial vertex and the final vertex. In citing function notation the initial vertex is xi where $T^{-1}x_i = \emptyset$. There is no other vertex in the network for which this is true. For the final vertex, xf, we have $Txf = \emptyset$. Again., this is the only vertex in the network for which this is true. For all other vertices $T^{-1}x \neq \emptyset$. The following section is an attempt to find the characteristics of PERT diagram that will be useful in analyzing citation index networks.

Using a citation network file, when given two papers, xi and xf , we would like to generate the network which relates them. From our point of view we can consider that theoretically we have a universal graph residing in a computer system. The problem then becomes to generate that portion of the entire graph which contains xi and xf so that the most meaningful papers are related to the given papers. In other words, given two vertices generate the graph which contains both of them. There are many such graphs for which this is true. Let us first consider two extreme cases. If, for example, we generate the entire graph in the computer, this would be impossible to handle as a search product since it would be analogous to asking a rather mundane question and getting all the world's knowledge for an answer. While this may not necessarily be the wrong answer it is certainly many magnitudes more than we would ordinarily want.

At the other extreme we have the minimum graph which contains xi and xf . To be precise, this would be nothing more than the two vertices. This of course would be of little value other than to verify whether or not xi and xf are in the graph. The problem at this extremum becomes more meaningful if we ask for the minimum connected graph which contains xi and xf. Indeed, one purpose of this section is to indicate how this graph can be defined. Associated with xi and xf are two times ti and tf respectivelv, where ti $\leq$ tf. We may think of xi therefore as being the vertex for which $T^{-1}x = \emptyset$ and we can think of xf as the vertex for which $Tx = \emptyset$. If these constraints do not permit the generation of a network we can lessen this requirement at a later time. (It is possible that a network would not result from this process.) Then we generate Tx_i; if xf ε Txi, then the procedure is completed. If not, then we generate

[1] J. E. Kelley, Jr., "The Critical Path Method: Resources Planning and Scheduling, " in Industrial Scheduling, J. F. Muth and G. L. Thompson, Eds., Prentice-Hall, Englewood Cliffs, N. J. , ch. 21 (1963).

[2] D. R. Fulkerson, "Expected Critical Path Lengths in PERT Networks, " Operations Research, 10(6): 808-817 (1962).

Txf and see if there exists an x such that x $\varepsilon\ T^{-1}x_f$ and x ε Tx x_i, that is $|T^{-1}x_f \cap Tx_i| \neq \emptyset$. If this is not true then we continue to generate a new set T^{-2}xi and go through the same process. We then go back and eliminate those vertices for which $T^{-1}x = \emptyset$ in the connected graph. And also eliminate all vertices where Tx = $\emptyset$. We have in the resultant graph a path with xi as its initial point and xf as its final point. This path can be called the history of xf with respect to xi which we call the critical historical path. Thus for our purposes, if any of the vertices along this path did not exist the relation between xf and xi would not exist. Let us now call the set of vertices in this path X critical, or more briefly, Xcr where any one of these elements is represented by x_c. It can be said that the research reported in every x_c depends on $T^{-1}x_c$. In order to include more vertices in the history, we might permit x to be added to the graph if some ancestor of x is x_c. Clearly $T^{-1}x_c \varepsilon X_c$, however, we would also have the case y = $T^{-1}x_c$ where $T^{-1}y \varepsilon X_c$. There may exist a path with n y's between any two xc's, however, the search for additional papers is completed when the t_y associated with each y is less than the tf associated with xf.

Procedure for Generating Critical Historical Network

Given x_1 and xf where $t_i \leq t_f$

1 GENERATE $T^{-m}x_i$ from m = 1 to m = k until xf $\varepsilon\ T^{-k}x_i$

2 IDENTIFY Xcr = $\{xi, x_1, x_2, x_3, x_f\}$

3 ELIMINATE al 1 x such as that Tx = $\emptyset$ or $T^{-1}x = \emptyset$ and repeat until all such x have been eliminated.

The resultant graph will be a PERT diagram in the true sense, and if we assign a value of one to each edge, X_{cr} will be the critical path of the PERT diagram.

Many other procedures could be hypothesized. However, among the procedured which are presented above or could be presented, it remains to be shown which would prove most valuable in an experimental situation.

CHAPTER VI

TRANSFORMATION GRAPHS

Given a citation index network or graph let us modify, in a rigorously prescribed manner, the initial graph so as to create another graph which has a meaningful relation to the initial graph. In mathematical terms, this is usually called a transformation. The resultant graph, the transformation graph, will ordinarily reveal properties which were obscured in the original graph.

The purpose of the following sections is to show that transformation graphs which result from operating on a citation index network can yield meaningful results, and thereby reinforce our initial assumption that graph theory is valuable for the analysis of citation index structures.

Author Analysis

A partitioning process can be useful in generating many kinds of graphs which are transformations of the original graph. One of the more interesting transformations shall be called the author influence graph. Consider the set of all vertices in the graph as the universe of discourse. We now perform a partitioning process so that each subset represents all the papers of a given author. Assume for this discussion that all papers are written by one and only one author. Thus we can create a transformation graph, H, where H = (X',T) where there is one vertex x' in H for each subset in G. Moreover, for every pair of vertices in G such that $y \varepsilon Tx$ we have a corresponding T-relation in H. This is illustrated graphically in Fig. 6.1

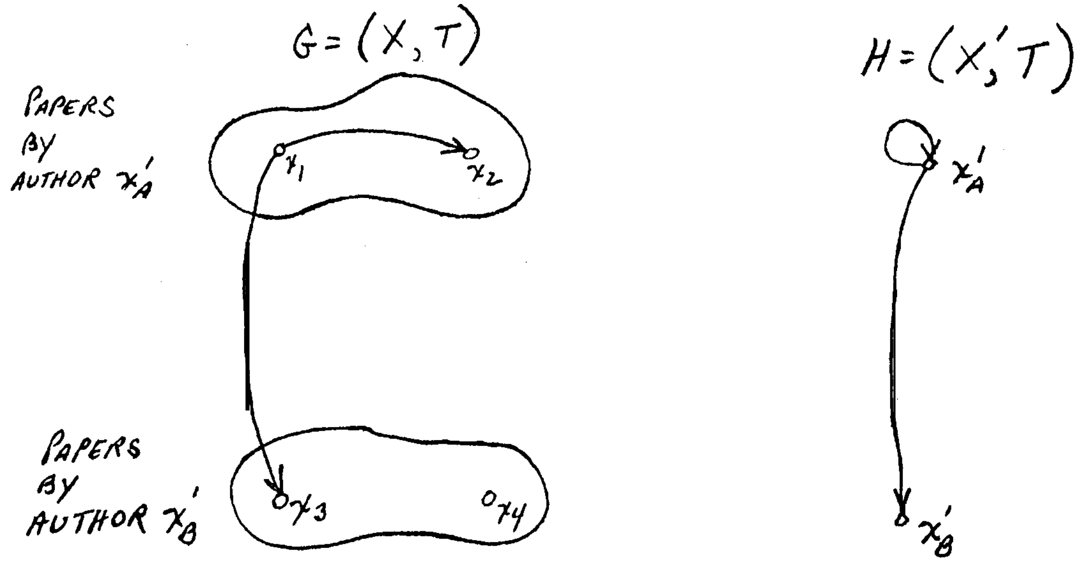


Fig. 6.1 – Sample Graph and its Transformation Graph

We see that x_1, x2 are papers which belong to the subset x'_A, which means x_1 and $x_2 \varepsilon x'_A$. Similarly, x3 and x4 are members of the subset x'_B, that is, x3 and $x_4 \varepsilon x'_B$. The citing function, T, can relate members of G in several ways. Firstly, we have $y \varepsilon Tx$ where y and x are both members of the same subset. This is called an interior edge. In Fig. 6.1, $x_1 \varepsilon T x_2$, where x_1 & x2& x'_A. Thus an interior edge is a definition of a self-citation. In H, we represent, this by putting a loop on x'_A. If there is more than one case of y ε Tx where (y , x) ε x_A , then we attach a number to the loop in H. This is the number of self-citations in G. Similarly, this is done for all authors which have self-citations, that is, this is done for all subsets which have interior edges.

The second type of citation is y ε Tx where y and x are in different subsets of the partitioning of G. In Fig. 6.1, $x_1 \varepsilon T x_3$ where $x_1 \varepsilon x'_A$ and $x_3 \varepsilon x'_B$ and $x'_A \cap x'_B = \emptyset$. Such an edge is called an edge of attachment. An edge of attachment represents a citation from one author to a different author. Each edge of attachment in G will have a corresponding edge in H. Their correspondence, however, will not ordinarily be one-to-one. By this it is meant that if there were more than one citation from x to y, the number of citations will be indicated in H, by a number associated with the respective edge.

Fig. 6.2 a-b are examples of hypothetical graphs which have resulted from an author transformation. These graphs illustrate the “degree of author interaction. "

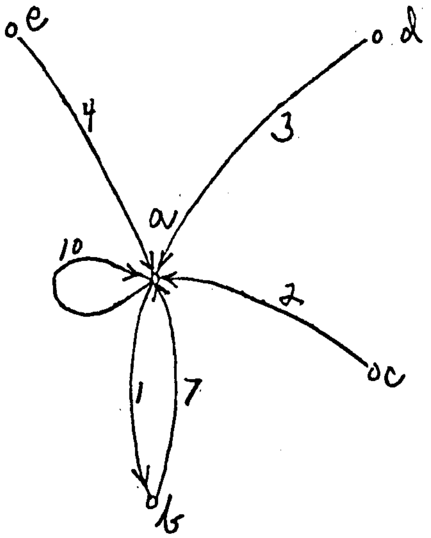


Fig. 6.2a -- Author Graph

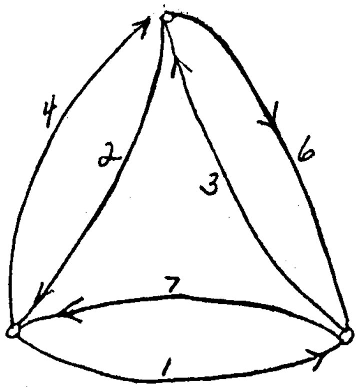


Fig. 6.2b -- Author Graph

In Fig. 6.2 a author a is the hub of the interaction. He is the most cited and does not cite c, d, and e. He recognizes the work of b, but b’s work exhibits a greater degree of dependency on a’s work than vice versa. The author graph in Fig. 6.2 a is significantly different from the author graph in Fig. 6.2 b. In the latter we find a high degree of author interaction among each other. This is a quantitative means for identifying invisible colleges. In the study of that aspect of sociology called group structure, this notion is called group cohesiveness.[1]

Now that the procedure for generating author transformation graphs is established, one can take two general approaches for further analysis. There are: (i) the discussion of the way the authors are related, i. e. , the connectivity, and (ii), the degree of connectivity, i. e. , the intensity of each relationship.

[1] Flament, op. cit., p. 32.

Types of Connectivity

If we have an author transformation graph H, we can identify at least four basic types of connectivity. The definitions to be presented here are for the most part a result of the work of Harary[2] and Luce.[3] Assume H = (X, T). Strongly connected means that for every pair of authors (x, y) of the graph H, there exists at least one path γ (xy). The most extreme case of strongly connectedness is the complete graph where every author is directly connected to every other author. The identification of a complete subgroup of an author transformation graph, is the identification of the nucleus of an invisible college. Semi-strongly connected is a graph such that either $\gamma(xy)$ or $\gamma(yx)$ exist. H is quasi-strongly connected if there exists a point z and a point z' such that $\gamma(xz)$, $\gamma(z'x)$, $\gamma(yz)$ and $\gamma(z'y)$ exist. Simply connected or weakly connected applies if there exists in H two authors such that there exists an undirected path (xy).

The above is an order of connectivity which is applicable to the analysis of the cohesiveness of the author transformation graph, H. This method is useful in describing the interaction of authors. For example, given any author, we can now find all other authors with whom he interacts at each level of cohesiveness or connectivity. The value of this method is clear when we realize one need only identify one key author in a field, and it is a relatively straightforward matter to generate a list of all other authors in the field by degree of activity in the field.

In general, the citation index network graph when operated upon by an appropriate partition process generates the author transformation graph. It remains, therefore, to identify the various subgraphs of H, which have the types ot connectivity discussed above. There has been some work done in other applications in this regard in the development of algorithms for locating graphs of various connectivities. In passing, we cite Roy's Algorithm 1961, which can be used to find the maximal strongly connected component of a graph by marking.[4]

Degree of Connectivity

The other aspect of the cohesiveness of the graph H, is the degree of connectivity. The removal of one edge from the graph may create one of the following effects on the graph: (1) it may raise the connectivity; (2) it may lower the connectivity or (3) it may leave the connectivity of the graph unaltered. To explain, in Fig. 6.3 a we have a subgraph of H, which indicates the direction of influence of one author on another. The number associated with each arc is the number of times a citation has taken place. If we like we could summarize the relationships between authors as in the graph in Fig. 6.3 b. Examining Fig. 6.3 b we see that we would have to remove three edges before the connectivity of the graph would be affected. Luce[5] would say, therefore, that the order of connectivity is 3. In Fig. 6.3 a, however, the notion of symmetry plays a role in determining connectivity; thus the removal of one edge in each direction would destroy the symmetry but not the connectivity.

[2] F. Harary, "A Criterion for Unanimity in French's Theory of Social Power, " in Studies in Social Power, Edited by D. Cartwright, Ann Arbor, Institute for Social Research (1959).

[3] R. D. Luce, "Two Decomposition Theorems for a Class of Finite Oriented Graphs," American Journal of Mathematics, 74(4): 701-722 (1952).

[4] Flament, op. cit., p. 35.

[5] R. D. Luce, "Connectivity and Generalized Cliques in Sociometric Group Structure, Psvchometrika, 15(2): 169-190 (1950).

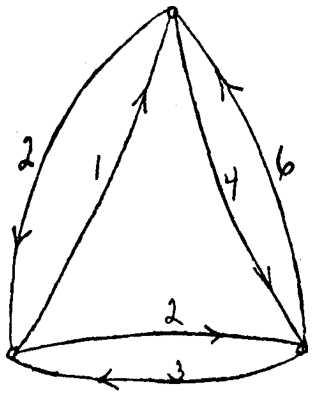


Fig. 6.3a -- Directed Author Graph

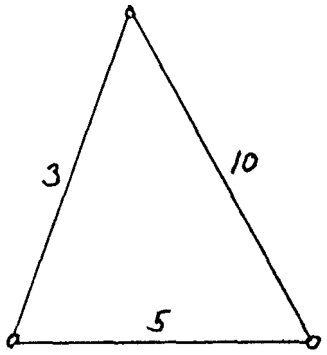


Fig. 6.3b -- Undirected Author Graph

The removal of an author may change the connectivity in the resultant graph(s). For example, in Fig. 6.4, author a is in a unique position in that if we remove him, the resultant graph is unconnected to say nothing of the effect on the connectivity. Such a vertex is called the articulation point. Ross and Harary[6] have presented methods for the identification of the articulation points.

We may say that if the removal of an articulation point results in two or more subgraphs, and if the resultant connectivities have raised or remained the same, we may interpret this to mean that the articulation point represents an author who is interdisciplinary in his research. Moreover, we have identified the authors of the other two or more fields in which his research is engaged.

Since many vertices will have edges going to and from other vertices, we can extend the concept of connectivity to embrace circuit connected graphs. If y = tx and (x, y) belongs to a circuit[7], then (x, y) is a circuit edge. If vertices (x_0, y_0) can be reached along a sequence of circuit edges, we say that x_0 and y_0 are circuit edge connected. The totality of vertices that are circuit edge connected with respect to x_0 is

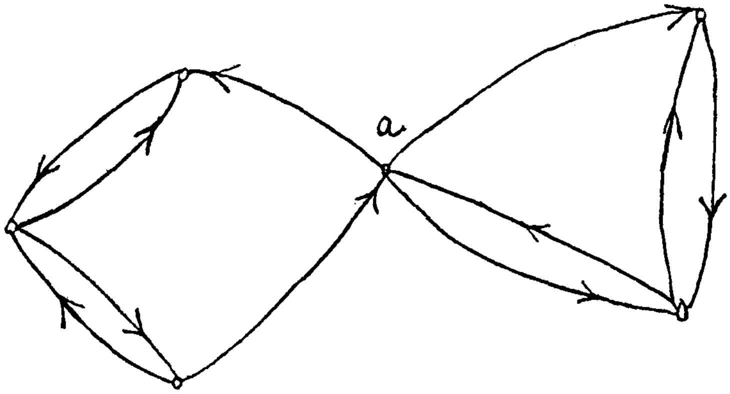


Fig. 6.4 -- Illustration of Articulation Point

[6] I. C. Ross and F. Harary, "Identification of the Liaison Persons of an Organization Using the Structure Matrix, " Management Science, 1(4):251-258 (1955).

[7] A circuit is a route which goes through no vertex more than once, and which returns to its starting point, i. e . , revisits only the beginning vertex.

called the leaf of x,, $L(x_0)$. The section graph of $L(x_0)$ is called the graph of the leaf L. We can further consider the concept of strongly circuit connected defined as two edges where there exists a sequence of circuits, where any pair of consecutive circuits have at least one edge in common. The vertex set of all strongly circuit connected to edge (xy) is called the lobe defined by (x, y) . Fig. 6.5 is a subgraph of the graph H. It is the author **transformation** graph. The arrowheads indicating the symmetry of the graph have been omitted for simplicity. The entire graph is a leaf graph with respect to every point since every vertex in the graph is circuit connected to every other point. Several pairs of authors in Fig. 6.5 form a lobe graph, for example, (c,d), (a, b), b, e), (e,f), etc. Clearly the leaf graph is composed of a number of lobes.

The following interpretation can be given to these concepts. The lobes define the intercitational relationships between authors and indicate how the work of authors depends on or relates to the work of other authors. The symmetry of the lobe can be meaningful in identifying authors whose work is closely related. It follows, therefore, that the composite of such authors constitute the leaf graph, which is in a sense the definition of an area of research. The formation of leaves, therefore, is a natural, pragmatic definition of an area of knowledge. One might say that the leaf is a branch of knowledge.

In general, we have provided a method such that given a large citation network we can generate a map of knowledge.

As an afterthought, one may question the assumption made at the outset that each paper was written by one and only one author. This need not be necessary if we do the following. In the network we will attribute to each author all the entries of the joint bibliography, and the joint citations. Moreover, we shall supply a citation to and from each of the joint authors to each other. This should sufficiently resolve this problem so as not to cause anyone any undue alarm.

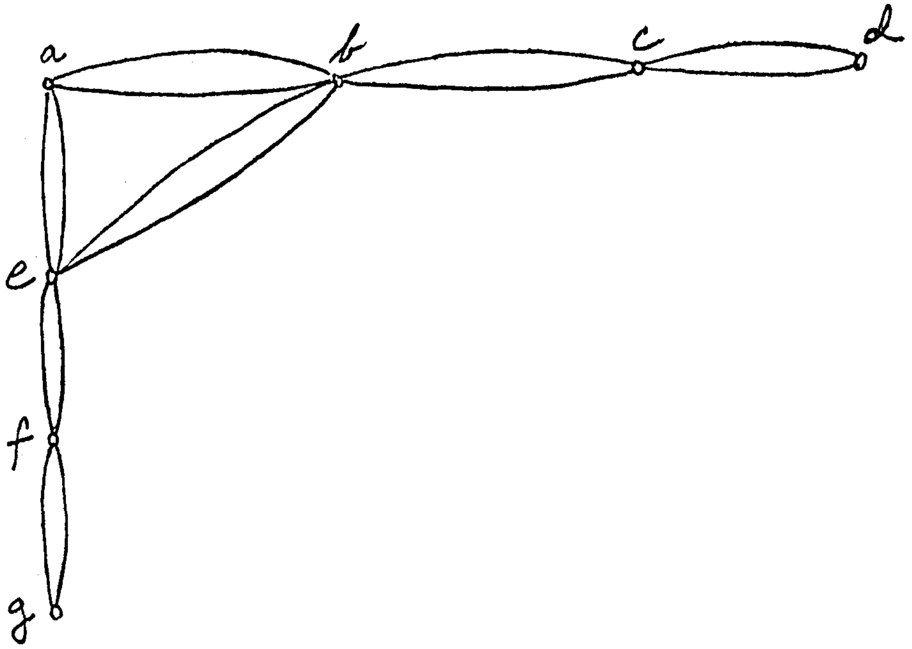


Fig. 6.5 -- Leaf Graph Containing Lobes

CHAPTER VII

CONCLUSION

The work has shown that the notation which is used in graph theory is applicable to the analysis of citation index structures. This notation was applied to all the aspects of citation indexing discussed in the literature. The value of the notation is its ability to make clear otherwise complicated statements, and thereby to force one to make rigorous statements when discussing citation indexing. Summarized below are the results of the application of this notation. Many of the concepts were hitherto obscured by the imprecision and ambiguous verbosity of natural language.

The graph theory notation has been used to define the citing function which permits me to uniquely describe the structure of a citation index network graph and its parts. The citing function has the power of relating groups of papers which are complexly connected through their bibliographies. The citing function permits these connection patterns to be explicitly stated. The use of the notation, for example, focuses attention on the difference between a bibliographic entry and a unique description of a scientific document. It forces one to realize the necessity of each document having a unique description. For a citation index system to function properly there must be established at the outset a method for entering items into the system so that each item is always described the same way no matter how many times it is entered into the system .

The use of an association matrix to represent a graph is a convenient tool for working with vast quantities of citation data. The standard matrix techniques are valuable for describing the characteristics of a citation index network. It permits one, for example, to identify quickly pertinent paths in the network.

When searching the citation index network, the use of graph theory gives one a tool for resolving the problems of both a batching search system and what has been called the “man-computer citation index” system. The graph theory notation facilitates the use of rigorous descriptions of cycling and bibliographic coupling, thus illustrating their differences and similarities, and indicating how each can be used in a citation index retrieval system. The successive application of the citing function is powerful in its ability to express any portion of the network structure. The flexible use of the citing function permits one to specify all types of relationships among vertices in a citation network, and thereby not be limited to a specialized case, e.g. , bibliographic coupling.

The critical path concept of a citation index network has been further expanded indicating the relationships among authors in a socio-historical milieu. The ability to transform a citation index network into an author influence graph facilitates by the use of graph theory, the analysis of the interaction of authors. It permits one to identify, for example, which authors constitute an invisible college. Moreover, it gives one a quantitative measure of the cohesiveness of the group of authors.

The use of graph theory is valuable in its ability to describe citation index structures. Moreover, it is of significant value in the formulation of the citation index search query. It is facile in its ability to specify transformation graphs which are useful for analyzing the interaction of authors. This work has, therefore, provided a mathematical description of citation indexing which permits the analysis of citation data so that many of the problems associated with citation indexing can be clearly stated and consequently resolved. It is hoped that this will result in non-ambiguous communication among researcher, system designer, programmer, user, and anyone else who finds it necessary to discuss citation indexing.

Area for Further Investigation

There are two main areas where further investigation would be worthwhile. The first would be to continue the mathematical presentation which has been introduced in this paper. This would constitute an attempt to discover if any additional graph theory techniques exist which are applicable to citation index analysis.

The second approach would be to do an analysis of the various types of cycling configurations to determine which are the most significant in terms of optimizing the relevance of the retrieved documents. This would most likely take the form of a statistical analysis of many cycling searches performed on a large-scale citation index file.

APPENDIX

TABLE 4

EXAMPLES OF THE USE OF THE CITATION FUNCTION

Citing Function Notation	English Equivalent
Tx_o	All papers which cite paper x_o
$T^{-1}x_o$	All papers in the bibliography of xo
$y_o \varepsilon T^{-1}x_o$	Paper y_o is a member of the bibliography of xo
$\left\| T^{-1}x_o \right\|$	The number of papers in the bibliography of xo
$y = T^{-1}(Tx_o)$	y is the set of all papers in the bibliographies of those papers which cite paper xo
$T^2x_1 = T\ (Tx_1)$	The set of papers which have cited the papers which have cited x_1
$Tx_o \cup T^2x_o$	All the papers which have cited x_o or have cited papers which have cited xo
Tx_1 n $T^{-1}x_2$	The set of papers which cite x_1 and are in the bibliography of x_2
Tx_1 n $Tx_2 \|$	The number of papers citing x_1 which also cite x_2
$\left\| T^{-1}y_1 \cap T^{-1}y_2 \right\|$	The number of papers in the bibliography of y_1 which also appear in the bibliography of Y2

TABLE 5

EXPLANATION OF SYMBOLS USED

Symbol	Explanation	Page First Mentioned
$=$	"Is equal to”	
$\neq$	“Is not equal to”	
$=$	"Is equal to by definition”	
$\hat{x}$	x is the symbol to be used when referring to the members of a set in general.	
$\hat{x}\hat{y}$	Same as above except the set is now a set of ordered pairs.	
$\{\ \}$	Braces used to enclose the members of a set.	
ε	“Is a member of”	
$\wedge$	“Is restricted to”	
$\cup$	"Or"	
$\cap$	“And”	
$\emptyset$	The null set (A set with no members.)	
$\subseteq$	“Is contained in”	
$\bigcup_{k=1}^{n} G_A^k$	“The union of elements of G_A^k as k goes from 1 to n"	
$\bigcap_{k=1}^{n} G_A^k$	“The intersection of elements of G_A^k as k goes from 1 to n. "	

BIBLIOGRAPHY

CITATION INDEXING

1. Garfield, E ., "Citation Indexes for Science -- A New Dimension in Documentation Through Association of Ideas," Science, 122 (3156):108-111 (1955).

2. ,________ "Citation Indexes In Sociological and Historical Research, " American Documentation, 14(4):289-291 (1963).

3. ________, "Science Citation Index -- A New Dimension In Indexing, " Science, 144(3619): 649-654 (1964).

4. ________, "Citation Indexing: A Natural Science Literature Retrieval System for the Social Sciences, " The American Behavioral Scientist, 7(10):58-61 (1964).

5. Kessler, M. M., "An Experimental Study of Bibliographic Coupling Between Technical Papers, " IEEE Trans. Information Theory IT9: 49-51 (1963).

GRAPH THEORY AND ITS APPLICATIONS

6. Abraham, C. T., "Graph Theoretic Techniques for the Organization of Linked Data, " Final report on Contract AF 19(628)-2752 ARCRL Cambridge, Mass. (1963).

7. ________, "Techniques for Thesaurus Organization and Evaluation, " Proceedings of the American Documentation Institute, p. 485-497 (1964).

8. Avondo-Bondino, G . , Economic Applications of the Theory of Graphs, Gordon and Breach, New York (1962).

9. Bedrosian, S. D., "Generating Formulas for the Number of Trees in a Graph, " Journal of the Franklin Institute, 277(4):313-326 (1964).

10. Beineke, L. W., and F . Harary, "On the Thickness of the Complete Graph," Bulletin of the American Mathematical Society, 70(4): 618-620 (1964).

11. Berge, C . , The Theory of Graphs and Its Applications, John Wiley and Sons, Inc., New York (1962).

12. Blicksman, S., "On the Representation and Enumeration of Trees, " Proceedings of the Cambridge Philosophical Society, 59(3):509-517 (1963).

13. Cartwright, D. and F. Harary, "Structural Balance, " Psychological Review' 63(5):277-293 (1957).

14. Coxeter, H. S. M., "Map-Coloring Problems, " Scripta Mathematica, 23: 11-25 (1957).

15. Dailey, C . A., "Graph Theory in the Analysis of Personal Document, " Human Relations, 12(1):65-74 (1959).

16. Dantzig, G. B., “On the Shortest Route Through a Network, " Report P. -134, Rand Corp. , Santa Monica, Calif.,(1959).

17. Dira, G. A., “A Property of 4-Chromatic Graphs and Some Remarks on Critical Graphs, " Journal of the London Mathematical Society, 27:85-92 (1952).

18. Doyle, L. B., "Sematic Road Maps for Literature Searchers, " Journal of the Association for Computing Machinery, 8(4):553-578 (1961).

19. Euler, L., “The Konigsberg Bridge Problem,” Scientific American 189 (1): 66-70 (1953).

20. Flament, C . , “Nombre de Cycles Complets dans un Réseau de Communication, " Bulletin du Centre d'Études et Recherches Psychotechniques, 8:105-110 (1959).

21. ________, Applications of Graph Theory to Group Structures, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1963).

22. Ford, G. W. and G. E. Uhlenbeck, “Combinatorial Problems in the Theory of Graphs, " I-IV, Proceedings of the National Academy of Sciences, 42: 122-128, 203-208, 529-535, 43: 163-167 (1957).

23. Ford, L. R. and D. R. Fulkerson, “Maximal Flow Through a Network, " Report I?.-605, Rand Corp., Santa Monica, Calif. (1954).

24. ________, Flows in Networks, Princeton University Press (1962).

25. Frucht, R., “Graphs of Degree Three with a Given Abstract Group, " Canadian Journal of Mathematics, 1(4):365-378 (1949).

26. Fulkerson, D. R., “Expected Critical Path Lengths in PERT Networks, " Operations Research, 10(6):808-817 (1962).

27. Glicksman, S., “On the Representation and Enumeration of Trees, " Proceed. Cambridge Philosophical Soc., 59(3):509-517 (1963).

28. Hakimi, S. L., “Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph, " Operations Research 12(3):450-459 (1964).

29. Harary, F., “On the Notion of Balance of a Signed Graph, " Michigan Mathematical Journal 2(2): 143-146 (1953).

30. ,________ “A Criterion for Unanimity in French’s Theory of Social Power, " in Studies in Social Power, D. Cartwright, Ed., AM Arbor, Institute for Social Reserach (1959).

31. ________, “A Graph Theoretic Method for the Complete Reduction of a Matrix with a View toward Finding its Eigenvalues, " Journal of Mathematics and Physics, 38(2):104-111 (1959).

32. ________, “Graph Theoretic Methods in the Management Sciences, " Management Science 5:387-401 (1959).

33. ________, "Graph Theory and Electric Networks, IRE Transactions on Circuit Theory, CT-6:95-109 (1959).

34. ________, "The Determinant of the Adjacency Matrix of a Graph, " SIAM Review, 4(3):202-210 (1962).

35. Harary , F . , G. Prins and W. T. Tutte, "The Number of Plane Trees, " Koninklij ke Nederlandse Akademie van Wetenschappen Proceedings, Series A Mathematical Sciences 6'7 (3):319-329 (1964).

36. Harary, F. and I. C. Ross, "The Number of Complete Cycles in a Communication Network, " J. of Social Psychology, 40(4):329-332 (1954).

37. ________, "A Procedure for Clique Detection Using the Group Matrix, "Sociometry 20(3):205-215 (1957).

38. Harrah, D . , Communication: A Logical Model, MIT Press, Cambridge (1963).

39. Hillman, D. J., "On Concept-Formation and Relevance, " Proceedings of the American Documentation Institute, p. 23-29 (1964).

40. ________ "Study of Theories and Models of Information Storage and Retrieval: Graphs and Algorithms for Term Relations, " Report 7, Center for the Information Sciences, Lehigh University (1964).

41. ________ "Mathematical Theories of Relevance with Respect to the Problems of Indexing: The Formal Basis of Relevance Judgments, " Report 1, Center for the Information Sciences, Lehigh University (1964).

42. Hu, T. C., "Multi-Commodity Network Flows, " Operations Research 11(3): 344-360 (1963).

43. Iverson, K. E . , "A Programming Notation for Trees, " Report RC-390, IBM Research Center, Yorktown, New York (1961).

44. ________, A Programming Language, John Wiley and Sons, Inc., New York (1962).

45. Kahn, A. B., "Topological Sorting of Large Networks, " Communications of the Association for Computing Machinery, 5(11):558-562 (1962).

46. Kalish, H. M., "Machine-Aided Preparation of Electrical Diagrams, " Bell Laboratory Record, 41(9): 338-345 (1963).

47. Kately, J . , "Automorphis Groups of Graphs, " Ph. D . dissertation, Michigan State University (1963).

48. Kelley, Jr., J. E., "Critical Path Planning and Scheduling: Mathematical Basis, " Operations Research 9(3):296-320 (1961).

49. ,________ "The Critical Path Method: Resources Planning and Scheduling, " in Industrial Scheduling, J . F . Muth and G. L. Thompson Eds . , Prentice-Hall, Englewood Cliffs, N. J., ch. 21 (1963).

50. Kochen, M., "Some Problems in Information Science with Emphasis on Adaptation to use Through Man-Machine Interaction, " Report AF 19 (628)-2752, Watson Research Center, Yorktown Heights (1964).

51. Kruskal, J . B., "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem, " Proc. Amer. Math. Soc. 7(1):48-50 (1956).

52. Lasser, D. J . , "Topological Ordering of a List of Randomly-Numbered Elements of a Network," Communications of the Association of Computing Machinery, 4(4): 167-168 (1961).

53. Luce, R. D., "Connectivity and Generalized Cliques in Sociometric Group Structure, " Psgchometrika 15(2): 169-190 (1950).

54. ________, "Two Decomposition Theorems for a Class of Finite Oriented Graphs," American Journal of Mathematics 74(4): 701-722 (1952).

55. _____ , "Networks Satisfying Minimality Conditions, " American Journal of Mathematics 75(11): 825-835 (1953).

56. Mayeda, W., "Properties of Classes of Paths, " Report R-212, Coordinated Science Laboratory, University of Illinois (1964).

57. Opler, A., "A Brief Survey of Topological Representations, " Proceedings of the American Documentation Institute, p. 499-502, (1964).

58. Ore, O., Theory of Graphs, American Mathematical Society, Providence, Rhode Island (1962).

59. ________ Graphs and Their Uses, Random House, New York (1964).

60. Otter, R., "The Number of Trees, " Annals of Mathematics 49(3):583-599 (1948).

61. Parker-Rhodes, A. F . , "On Talking to Computers, " Proceedings of the American Documentation Institute, p. 477-484 (1964).

62. Ramamoorthy, C . V. and D. W. Tufts, "Generating Functions of Abstract Graphs with Applications, " Report 439 on Contract NR-372-012, Harvard University, Cambridge (1964).

63. Reed, M. B . , Foundations for Electric Network Theory, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1961).

64. Reisner, P., "A Note on Minimizing Search and Storage of Theasaurus Network by Structural Reorganization of the Net, " Final Report on Contract AF 19(628)-2752 p. 365-373 (1964).

65. Rescigno, A. and G. Segre, "On Some Topological Properties of the Systems of Compartments, " Bulletin of Mathematical Biophysics, 26 (1): 31-38 (1964).

66. Robertson, N., "The Smallest Graph of Girth 5 and Valency 4, " Bulletin of the American Mathematical Society, 70(6): 824-825 (1964).

67. Rosenblatt, D., "On the Graphs and Asymptotic Forms of Finite Boolean Relation Matrices and Stochastic Matrices," Naval Research Logistics Quarterly 4(2):151-168 (1957).

68. ________, "On the Graphs of Finite Idempotent Boolean Relation Matrices," Journal of Research of the National Bureau of Standards--B, Mathematics and Mathematical Physics 67B(4): 249-256 (1963).

69. Ross, I. C. and F. Harary, "On the Determination of Redundancies in Sociometric Chains," Psychometrika 17(2): 195-208 (1952).

70. ________, "Identification of the Liaison Persons of an Organization Using the Structure Matrix," Management Science 1(4):251-258 (1955).

71. ________, "A Description of Strengthening and Weakening Members of a Group," Sociometry 22(2): 139-147 (1959).

72. Ryser, H. J., "Combinatorial Properties of Matrices of Zeros and Ones," Canadian J. Math 9(4):371-377 (1957).

73. Saaty, T. L., "The Minimum Number of Intersections in Complete Graphs," Proceedings of the National Academy of Sciences 52(3):688-690 (1964).

74. Salton G., "Information Storage and Retrieval," Report ISR-2, The Computation Laboratory, Harvard University (1962).

75. ________, "Manipulation of Trees in Information Retrieval," Communications of the Association for Computing Machinery, 5(2): 103-114 (1962).

76. Salton, G. and E. H. Sussenguth, Jr., "A New Efficient Structure-Matching Procedure and Its Application to Automatic Retrieval," Proceedings of the American Documentation Institute, p. 143-146 (1963).

77. Schurmann, A., "The Application of Graphs to the Analysis of Distribution of Loops in a Program," Information and Control, 7(3): 275-282 (1964).

78. Scidmore, A. K. and B. L. Weinberg, Storage and Search Properties of a Tree-Organized Memory System, Communications of the Association for (Comput1ng M9chinery6 6(1):3 28-31) .

79. Simon, A., Models of Man, John Wiley and Sons, Inc., New York (1957).

80. Singleton, R. R., "On Minimal Graphs of Maximum Even Girth," Ph.D. dissertation, Dept. of Mathematics, Princeton University (1962).

81. Sussenguth, Jr., E. H., "Structure Matching In Information Processing," Ph. D. dissertation, Department of Applied Mathematics, Harvard University (1964).

82. ____ , "A Graph-Theoretic Algorithm for Matching Chemical Structures," J. Chem. Documentation 5(1): 36-43 (1965).

83. Trauth, C. A., "On the Connectedness of Directed Graphs under Binary Operations," Ph.D. dissertation, Dept. of Mathematics, University of Michigan, Ann Arbor (1963).

84. Ungar, P., "On Diagrams Representing Maps, " J . London Math. Soc . , 28: 336-342 (1953).

85. Whyburn, G. T., "Generic and Related Mappings, " Bulletin of the American Mathematical Society 69(6):757-761 (1963).

86. Windley, P. F., "Trees, Forests and Rearranging, " The Computer Journal 3(1):84-88 (1960).

87. Wollmer, R . , 'Removing Arcs from a Network, " Operations Research 12(6): 934-940 (1964).

88. Youngs, J. W. T., "Simplest Imbeddings of the Complete 12 Graph, " Report P.-2426, RAND Corp., Santa Monica, Calif. (1961).

89. ________, "Remarks on the Genus of a Complete Graph, " Report P. -2428, RAND Corp., Santa Monica, Calif. (1961).

BACKGROUND MATHEMATICS

90. Bickley, W. G., and R . S. H. G. Thompson, Matrices, Their Meaning and Manipulation, D. Van Nostrand Company, Inc., Princeton, New Jersey (1964).

91. Birkhoff, G. and S . MacLane, A Survey of Modern Algebra, Rev. Ed., The Macmillan Company, New York (1953).

92. Feferman, S., The Number Systems, Foundations of Algebra and Analysis, Addison-Wesley Publishing Company, Inc., Reading, Mass. (1964).

93. Kamke, E., Theory of Sets, Dover Publications, New York (1950).

94. Quine, W. V. O., Mathematical Logic, Harvard University Press, Cambridge (1961).

95. Sawyer, W. W., A Concrete Approach to Abstract Algebra, W. H. Freeman and Company, San Francisco (1959).

96. Stoll, R . R . , Sets, Logic, and Axiomatic Theories, W. H. Freeman and Company, San Francisco (1961).

97. Suppes, P., and S. Hill, First Course in Mathematical Logic, Blaisdell Publishing Company, New York (1964).