Algorithmic Citation-Linked Historiography—Mapping the Literature of Science

Eugene Garfield
Chairman Emeritus, ISI, 3501 Market Street, Philadelphia, PA 19104.
Email: Garfield@codex.cis.upenn.edu

A. I. Pudovkin
Institute of Marine Biology, Russian Academy of Sciences, Vladivostok 690041, Russia.
Email: aipud@online.ru

V. S. Istomin
Center for Teaching, Learning, and Technology Washington State University, Pullman, WA 99164-4550.
Email: Vi@mail.wsu.edu

There is a large literature on mapping and visualizing the scholarly literature (White & McCain, 1997; Buter & Noyons, 2001). However, none of these methods have been used to create historical displays of works on a given subject. The authors have developed a process and software called *HistCite* for generating chronological maps of collections resulting from searching the *ISI Web of* Science (*WOS)*, SCI/SSCI/AHCI on CD-ROM or SciSearch on Dialog. Export files are created in which all cited references for source documents are captured. These files are processed by *HistCite* to generate tables of the most-cited works. Real time demonstrations of several topics such as bibliographic-coupling, co-citation analysis, gene flow, etc. will be provided. *The HistCite* software includes an expert system for detecting and editing errors or variations in cited references. Export Files of 1,000 or more records are processed in minutes on a PC. Ideally the system will be used to help the searcher quickly identify the most significant work on a topic and trace its year-by-year development.

Historical Introduction

Even before the advent of the *Science Citation Index* in print the use of citation data to help write the history of science and scholarship was discussed. In 1964 Garfield, Sher and Torpie issued their report on "The use of Citation Data in Writing the History of Science" (Garfield, Sher, & Torpie, 1964) which included an historiograph of DNA's history from Mendel to Nirenberg. Flow charts of the many papers on this topic were created manually based on the references cited in a set of core source documents. Then, in subsequent years Garfield periodically reported on the potential use of citation indexes for historiography (Garfield, 1971).

In characterizing literature searches on topics such as acoustics, etc., A. E. Cawkell used manually created historical maps (Cawkell, 1989, 2000). In Garfield's information retrieval course at the University of Pennsylvania Moore School of Electrical Engineering, students were required to create historiographs of topics of their own choosing. In all these mapping exercises it was explicitly assumed that scholars would use *ISI*'s citation indexes to obtain the records needed to create historical maps by manual methods. These historiographs or historiograms would aid in the study of the history of contemporary scientific topics. Since history and bibliography were intimately linked the term historio-bibliography was created (Garfield, 1971).

During the DNA History project, the idea of writing computer programs that would create such maps directly from the electronic files of the *Science Citation Index* was often discussed. This would require rapid random access to massive files that could retrieve both cited and citing documents in real time. In the 1960s, however, low cost gigabyte memories were still a dream. The implementation of real time mapping had to wait for the time when computer memories were large and cheap enough to handle retrospective files covering many decades of literature. While on-line searches were possible in the 1970s, mapping in real time was still not feasible because the PC had not yet come along. Only when the output of a completely linked large file of thousands of records could be handled by today's PCs did the creation of historiographs in real time become feasible.

There have been many different types of "mapping" exercises performed on a small scale particularly with respect to co-citation clustered files of bibliographic information. In the past, clustering required main frame computers (Small & Garfield, 1985; Garfield, 1992) and in most cases still does. These ideas were later extended to creating small cluster maps on-line as in the *SciMap* system developed by Henry Small at ISI. In that mainframe system a starting paper is used to seed the creation of a co-citation cluster map (Small, Sweeney, & Greenlee 1985; Small, 1994). In spite of the many mapping and visualization techniques available none of them were applied to the creation of historiographs. Indeed, none of the many authors on co-citation mapping considered the significant relationship between historical display and its potential role in evaluating the output of searches with *Science Citation Index* (*SCI*), *Medline, Chemical Abstracts,* etc.

From the earliest days until quite recently, we thought of creating historiographs by seeding one or two primordial papers. Then the *Science Citation Index* would be used to trace forward in time all the papers that had cited the starting references. This is the essence of doing what has now become a traditional cited reference search. Indeed, the basic purpose of an historiograph is to display the chronological development of a topic or field -- from the primordial paper forward, year by year. This notion was also influenced by the fact that the published *SCI* appeared annually both in print and on CD-ROM. And such searches could be amplified by cycling, that is, searching forward and backwards on cited references.

In contrast, traditional literature searches are focused on retrieving the most current material and then working backward. Our initial experiments involved the use of the annual CD-ROMs to perform a cited reference search on a single starting paper. All papers that cited it the first year were retrieved. Then a further cited reference search was done on those citing papers. The process was iterated for as many years of the literature as was necessary.

For each year of the literature searched there would be 0 to N papers retrieved. The full *SCI* Source record for each of the N papers would be captured including not only authors, titles, journal, volume page and year, but also their lists of cited references. If we assume that there are 20 cited references per source paper then there would be 20N cited references collected. The purpose of these multigenerational searches is to build up a file of relevant source documents together with a much larger file of cited references. Thus, if the collection involves 500 source papers, the list of uniquely cited references will be from 5,000 to 10,000 items.

HistCite Program

The *HistCite* program described herein sorts cited references to create a virtual mini-citation index. In addition, a variety of sort keys is used to create ranked directories of authors, journals, and citation frequency. The mini-citation index is sorted to determine the citation frequency within the collection for each uniquely cited document. If 1,000 different papers or books have been cited then the average frequency will be 10. However, the range of frequency would run from 1 to 1,000. The system also creates a series of tables or matrices which list all the 500 source documents, each of which is assigned a number from 0 to 500, 0 being the primordial paper in a citation-based search or the oldest in a keyword-based search. The table is arranged in chronological order.

Since the record for each source document contains its citation frequency, both within the local mini database and the global *SCI,* these data can also be used to create tables sorted by citation frequency. This identifies groups of core, high impact documents that are cited above an arbitrarily chosen threshold. If there are 500 source papers then a 5% selection threshold would produce 25 core papers. These core papers are of prime interest especially to a searcher who is not familiar with the subject matter. The core list contains those candidate papers one would examine first in reviewing the topic. The coordinates of these papers will also be used later to create an historiograph of the topic which displays each of the papers and their citation links chronologically.

Identifying Core Literatures

As stated earlier, it was initially assumed we would begin with one primordial paper. However, it became apparent that one could feed in groups of papers by one or more authors – and by extension, larger clusters of papers by institution or by key word. Thus the output of any conventional search or a combination of citation and key word searches could be input to the system. Once the input bibliography is created, the core papers on the "topic" can be identified.

Visualization

The production of the various tables or lists from these procedures is of course separate from the problem of visualizing these data in the form of maps or graphs. These artifacts aid in the visual perception of the interrelationships between citing and cited papers. Creating maps of related documents present problems in display due to the limitations of space and restrictions of the 8 x 11 piece of paper. Visualization is aided by using larger pages. However, the advent of computer display screens means that one can create a virtual display of unlimited size. In future versions of the software, segments of the chronological maps will be shown in a movable display. Using mouse clicks and pop-up windows one would first show a condensed version of a large map in which the main nodes are visible but intermediate nodes are not. Thus a map of several hundred nodal papers would first be seen in a condensed version in which only 25 to 50 nodes are seen, perhaps the most cited papers in the collection or the most-cited in a particular time period. The full map could be observed in chronological sections from top to bottom or from left to right. Essentially one goes from a standard two dimensional display to a moving interactive multi-dimensional display. The combination of computer with human selection permits the algorithmic real time visualization of the historical connections between documents on a micro or macro level and identifies the most significant nodes. Thus the user can quickly perceive the historical connections between the core documents. The system could also include classification tags or research front identifiers that would permit one to recognize a larger cluster or category of which each paper is a part. While the initial system is limited to processing source records containing index tags and abstracts, entire texts of documents or Internet links could also be included. In that way it would be possible to observe the contextual significance of each citation as illustrated by the work of Steve Lawrence (Lawrence, Giles, & Bollacker, 1999).

In our initial implementation of the *HistCite* system, records from the *Science Citation Index* have been used. The initial implementation produces numerous basic indexes based on year, author, journal, local and global citation frequency, and missing links. Other tables will be added in future versions as e. g. title or key word lists which identify the most used terminology.

Outer References

Significantly, the system also produces frequency ranked indexes of outer references, that is, cited papers and books that fall outside the starting collection. These are works which do not turn up in the original search but are cited in the collection. The user can examine these candidate references and decide whether to add them to the nodal group. For example, a highly cited book or patent might be cited which is not part of the original *SCI* source database. For each of these a source record would have to be created . Some of these items may in fact have been published prior to the starting reference. In the latest version of the software a module has been added which provides a semi-automatic look-up of the full source record for each outer reference that is included in the *SCI.* Similar hotlinks could be provided to other databases. Once the full record is found, it can be added to the original bibliography and the software invoked to create a new matrix or map. The size of the outer reference list can be specified by the user and may even run into the thousands.

Missing References

It is well known that authors cite references with many variant spellings or make errors in one or more parts of the reference such as volume or page. These "missing" references are identified in a separate table. As part of the procedures invoked, the program will seek out the closest matched document in the collection and suggest candidates to examine. This can be done manually or by an expert system which e.g. adds a missing volume number to a citation that is otherwise identical for author, journal, year and page. In a large number of references the page cited will not be the first page. In many chemical papers when a chemical compound is mentioned, the exact page is cited. If that page is cited often, the user may wish to treat that as a separate source. Otherwise, the page number could be changed or deleted, so that it will be included in the citation frequency score for the fully paginated reference. If it is a duplicate, it will be deleted when the program is run a second time.

To reiterate, an expert system facilitates the process of editing errors or variations so that they can be integrated with the main collection. Fortunately the number of such errors or variations is normally quite manageable. The reader can edit them or not. Most are singletons that ordinarily will not affect the overall ranking or mapping results.

All of the processes described with respect to the *SCI* can of course be extended to output data from other databases but in one way or another the source documents should include cited reference links. This could be done by matching non-*SCI* files with *SCI* data or by manually or electronically extracting cited references from the full texts of papers to which they are linked. *CAS, PscychInfo* and other data bases now provide links to full texts. The entire full text of over 100 years of *Science, Physical Review*, etc. is now available as are more recent years of thousands of journals on vendors such as HighWire, JSTOR, Ingenta, and elsewhere.

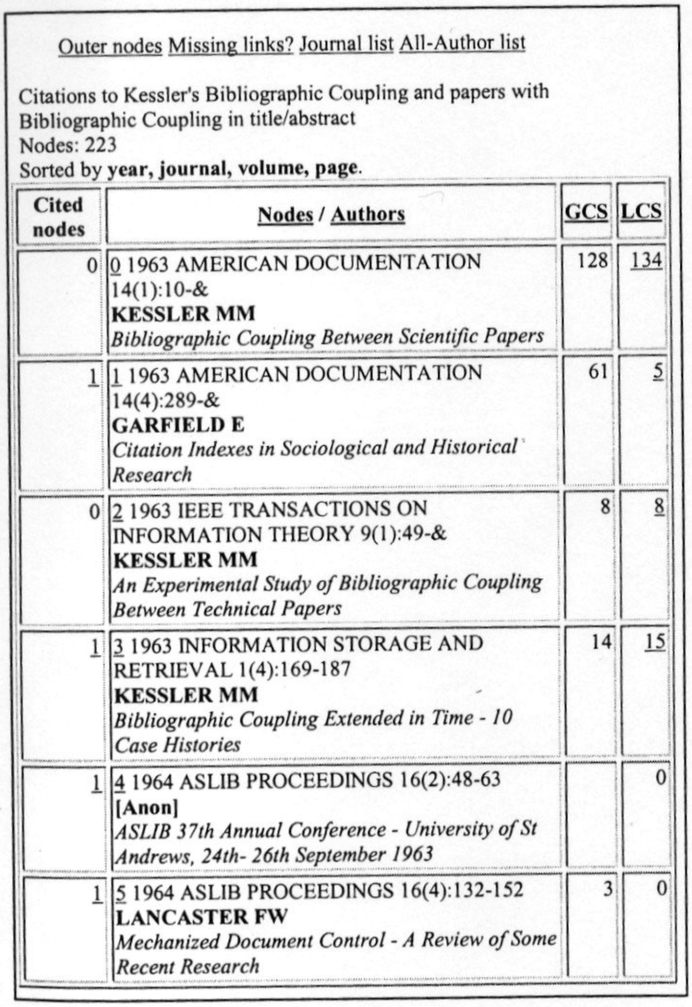
In all these cases it is necessary to standardize the formats of cited references so they can be consistently processed by the system. For that reason conversion programs can be included which take data in its original format and converts it to the required system format . At present, the program accepts export records from either the *SCI CD-ROM.* the *Web of Science* or SciSearch on Dialog. The basic coden of name, journal, volume, page and year is sufficient for most cited references. However, in the source records it is helpful to include the date or issue number to better separate papers chronologically especially for fast moving topics.

In the examples which follow the collection was created by searching the *Web of Science*. The files were obtained by doing either a title word or cited reference search. The resulting collections were downloaded in the WOS EXPORT format which is in plain ASCII text. The plain ASCII text file containing the collection is processed by the program. The output is then presented in html. The *HistCite* program can reside on the user's hard drive or a server.

HistCite Features

Chronological Table

The initial output of the system is the master table sorted in chronological order (see Figure 1). The first primordial document is labeled 0 and the remainder following from 1 to N. A click on the node number opens the complete source record.

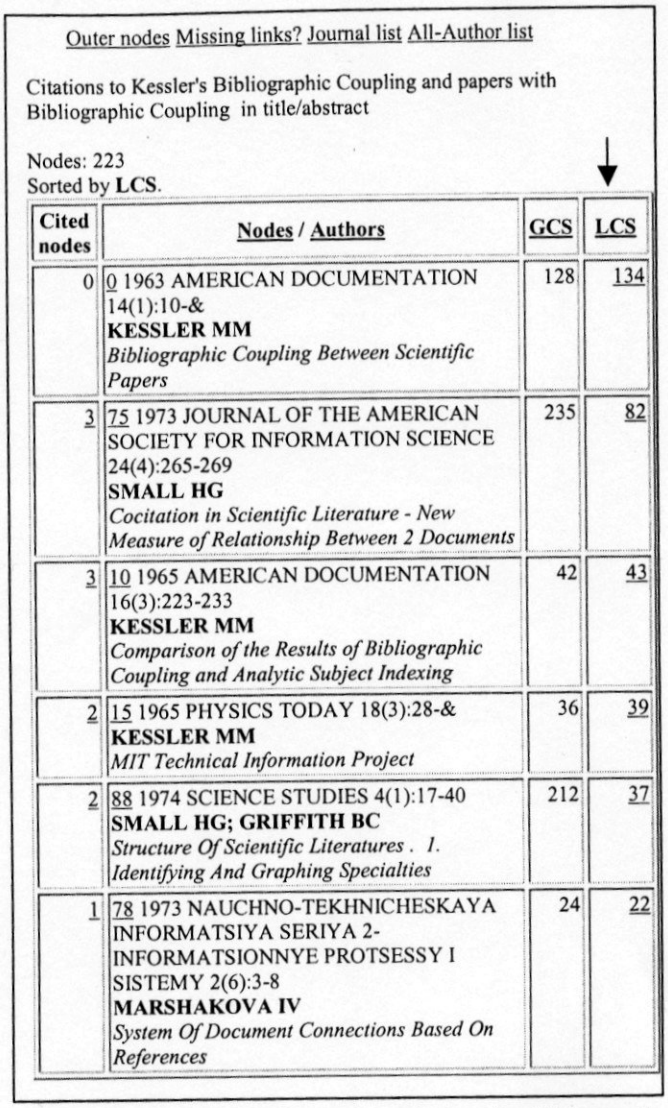


Outer nodes Missing links? Journal list All-Author list

Citations to Kessler's Bibliographic Coupling and papers with Bibliographic Coupling in title/abstract
Nodes: 223
Sorted by **year, journal, volume, page.**

Cited nodes	Nodes / Authors	GCS	LCS
0	0 1963 AMERICAN DOCUMENTATION 14(1):10-& **KESSLER MM** *Bibliographic Coupling Between Scientific Papers*	128	134
1	1 1963 AMERICAN DOCUMENTATION 14(4):289-& **GARFIELD E** *Citation Indexes in Sociological and Historical Research*	61	5
0	2 1963 IEEE TRANSACTIONS ON INFORMATION THEORY 9(1):49-& **KESSLER MM** *An Experimental Study of Bibliographic Coupling Between Technical Papers*	8	8
1	3 1963 INFORMATION STORAGE AND RETRIEVAL 1(4):169-187 **KESSLER MM** *Bibliographic Coupling Extended in Time - 10 Case Histories*	14	15
1	4 1964 ASLIB PROCEEDINGS 16(2):48-63 **[Anon]** *ASLIB 37th Annual Conference - University of St Andrews, 24th- 26th September 1963*		0
1	5 1964 ASLIB PROCEEDINGS 16(4):132-152 **LANCASTER FW** *Mechanized Document Control - A Review of Some Recent Research*	3	0

Figure 1: Chronological File of Papers Citing Kessler 1963



Outer nodes Missing links? Journal list All-Author list

Citations to Kessler's Bibliographic Coupling and papers with Bibliographic Coupling in title/abstract

Nodes: 223
Sorted by **LCS.**

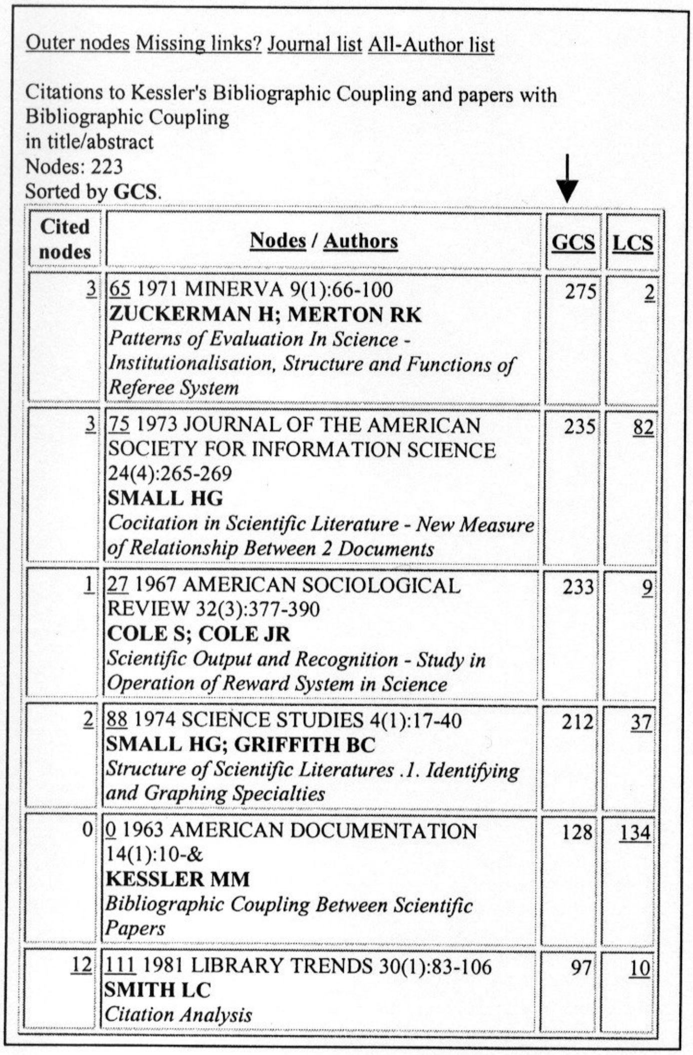
Cited nodes	Nodes / Authors	GCS	LCS
0	0 1963 AMERICAN DOCUMENTATION 14(1):10-& **KESSLER MM** *Bibliographic Coupling Between Scientific Papers*	128	134
3	75 1973 JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 24(4):265-269 **SMALL HG** *Cocitation in Scientific Literature - New Measure of Relationship Between 2 Documents*	235	82
3	10 1965 AMERICAN DOCUMENTATION 16(3):223-233 **KESSLER MM** *Comparison of the Results of Bibliographic Coupling and Analytic Subject Indexing*	42	43
2	15 1965 PHYSICS TODAY 18(3):28-& **KESSLER MM** *MIT Technical Information Project*	36	39
2	88 1974 SCIENCE STUDIES 4(1):17-40 **SMALL HG; GRIFFITH BC** *Structure Of Scientific Literatures . 1. Identifying And Graphing Specialties*	212	37
1	78 1973 NAUCHNO-TEKHNICHESKAYA INFORMATSIYA SERIYA 2-INFORMATSIONNYE PROTSESSY I SISTEMY 2(6):3-8 **MARSHAKOVA IV** *System Of Document Connections Based On References*	24	22

Figure 2: Most-cited papers from the starting bibliography ranked by local citation score (LCS)

In addition to the basic chronological "home" table several sorts can be called out. These are activated by clicking on the hot links. The first is the Local Citation Score (see Figure 2). This local score (LCS) is based on the citation frequency within the basic collection.

A second display, (GCS), sorts by the global citation frequency, that is, how often each paper is cited in the entire *SCI* (see Figure 3).

These sorts are quite useful in performing literature searches as the reader can view immediately the core papers on the topic, that is the highest impact papers.



Outer nodes Missing links? Journal list All-Author list

Citations to Kessler's Bibliographic Coupling and papers with Bibliographic Coupling
in title/abstract
Nodes: 223
Sorted by **GCS**.

Cited nodes	Nodes / Authors	GCS	LCS
3	65 1971 MINERVA 9(1):66-100 **ZUCKERMAN H; MERTON RK** *Patterns of Evaluation In Science - Institutionalisation, Structure and Functions of Referee System*	275	2
3	75 1973 JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE 24(4):265-269 **SMALL HG** *Cocitation in Scientific Literature - New Measure of Relationship Between 2 Documents*	235	82
1	27 1967 AMERICAN SOCIOLOGICAL REVIEW 32(3):377-390 **COLE S; COLE JR** *Scientific Output and Recognition - Study in Operation of Reward System in Science*	233	9
2	88 1974 SCIENCE STUDIES 4(1):17-40 **SMALL HG; GRIFFITH BC** *Structure of Scientific Literatures .1. Identifying and Graphing Specialties*	212	37
0	0 1963 AMERICAN DOCUMENTATION 14(1):10-& **KESSLER MM** *Bibliographic Coupling Between Scientific Papers*	128	134
12	111 1981 LIBRARY TRENDS 30(1):83-106 **SMITH LC** *Citation Analysis*	97	10

Figure 3: Papers that cite Kessler 1963 ranked by Global Citation Score (GCS)

All-Author Table

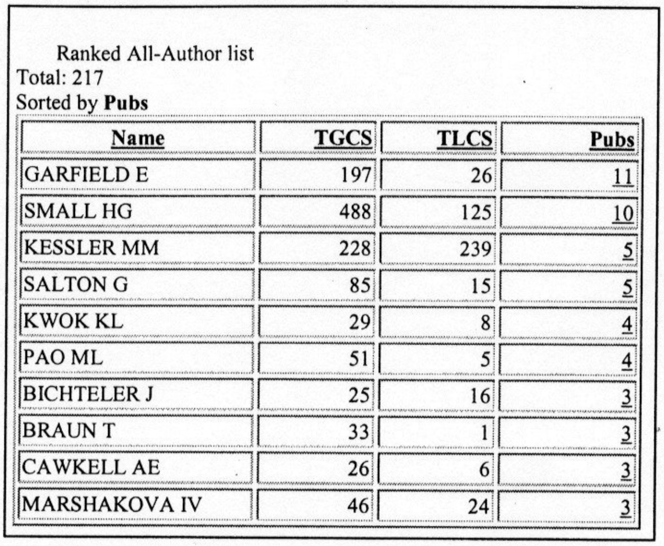
Figure 4 lists all authors in the collection. The most published author is at the top. Hotlinks permit the display of the authors by global or local citation frequency. Thus the most-cited authors, as distinguished from the most published authors can be shown. The individual citation frequencies for these papers are totaled.

Journal Table

In Figure 5, the journals in which the papers were published are displayed. At the right is shown the total number of papers for that journal.

Outer References

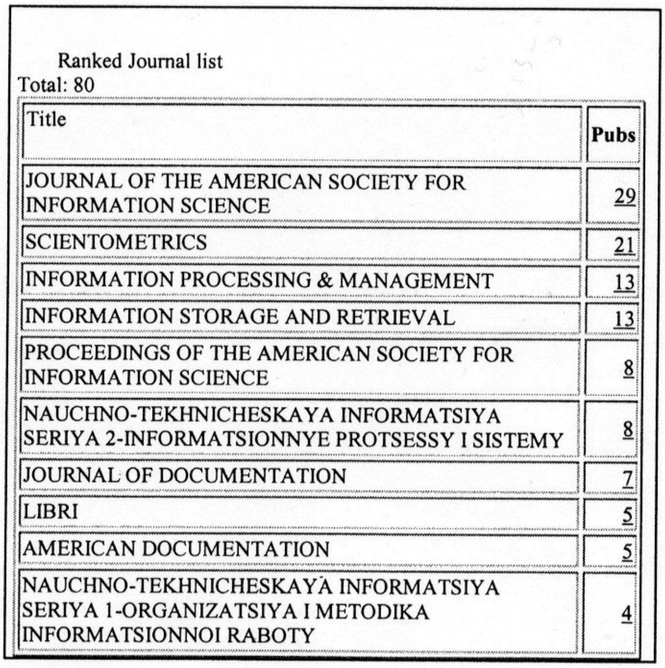
The "outer nodes" link lists the thousands of cited references that are not part of the basic collection or bibliography (see Figure 6). These "outer nodes" include books, patents, etc., as well as papers not included as sources in the *SCI*. They are sorted by citation frequency in the local network. The user can choose to retrieve such items and provide the missing source data for those that are cited above a given threshold. Once identified and expanded to include their list of cited references, full bibliographic data for these nodes can be added to the basic collection. The program is run again so that the new papers are integrated into the collection. A semi automatic look-up



Ranked All-Author list
Total: 217
Sorted by **Pubs**

Name	TGCS	TLCS	Pubs
GARFIELD E	197	26	11
SMALL HG	488	125	10
KESSLER MM	228	239	5
SALTON G	85	15	5
KWOK KL	29	8	4
PAO ML	51	5	4
BICHTELER J	25	16	3
BRAUN T	33	1	3
CAWKELL AE	26	6	3
MARSHAKOVA IV	46	24	3

Figure 4: Authors ranked by number of publications

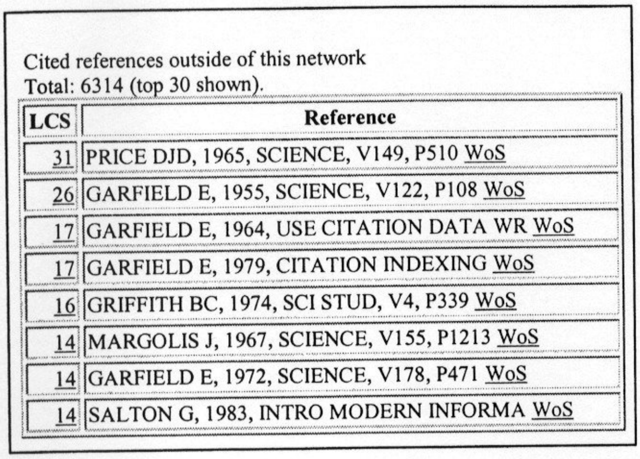


Ranked Journal list
Total: 80

Title	Pubs
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE	29
SCIENTOMETRICS	21
INFORMATION PROCESSING & MANAGEMENT	13
INFORMATION STORAGE AND RETRIEVAL	13
PROCEEDINGS OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE	8
NAUCHNO-TEKHNICHESKAYA INFORMATSIYA SERIYA 2-INFORMATSIONNYE PROTSESSY I SISTEMY	8
JOURNAL OF DOCUMENTATION	7
LIBRI	5
AMERICAN DOCUMENTATION	5
NAUCHNO-TEKHNICHESKAYA INFORMATSIYA SERIYA 1-ORGANIZATSIYA I METODIKA INFORMATSIONNOI RABOTY	4

Figure 5: Journals ranked by number of papers published

of each item can be activated if the user has access to *Web of Science* by clicking on WOS. These outer nodes can also be scanned with the FIND command to locate all variant citations for the same cited work .



Cited references outside of this network
Total: 6314 (top 30 shown).

LCS	Reference
31	PRICE DJD, 1965, SCIENCE, V149, P510 WoS
26	GARFIELD E, 1955, SCIENCE, V122, P108 WoS
17	GARFIELD E, 1964, USE CITATION DATA WR WoS
17	GARFIELD E, 1979, CITATION INDEXING WoS
16	GRIFFITH BC, 1974, SCI STUD, V4, P339 WoS
14	MARGOLIS J, 1967, SCIENCE, V155, P1213 WoS
14	GARFIELD E, 1972, SCIENCE, V178, P471 WoS
14	SALTON G, 1983, INTRO MODERN INFORMA WoS

Figure 6: Outer nodes – most cited works outside original bibliography

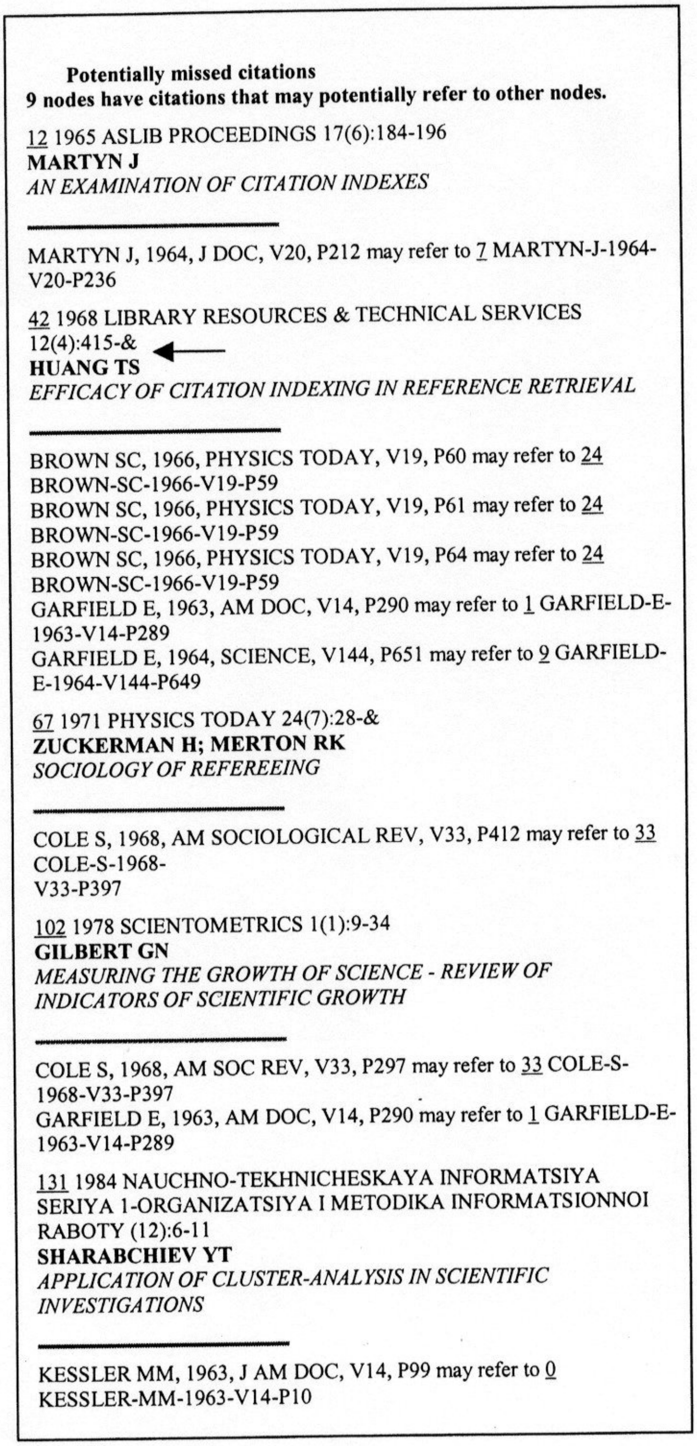
Missing Links

The system identifies questionable or "missing" citations where there is reason to believe there is an error or variation that prevents unification (see Figure 7). This expert system permits the reader to correct errors of omission in volume, number, year, pagination, etc. Once corrected these items can be fed back into the file so as to refine the data and complete the citation counts and maps. In the example shown for T. S. Huang's paper, he has cited specific page numbers in the same article by S. C. Brown.

For bibliographies up to 1,000 papers, the mapping procedure takes approximately one minute, that is, once the output bibliography is downloaded from WOS. Using the latest version of WOS approximately 150 records can be exported at one time. Thus to accommodate 1,000 records, one would have to run the search in seven or eight segments, depending upon the number of references cited . The various outputs are cut and pasted into one consolidated export file. *HistCite* eliminates any duplicates encountered.

Single Journal Output

We have tested the program on files as large as 5,000 papers. In one case, all papers were published in a single journal – *Evolution* (see Figure 8). In order to create a pure chronological display, the sort key has to include not only volume, page and year, but also issue number. As is well-known, the WOS output file is arranged in approximate reverse chronological order based on production input. The ISI production procedure involves the input of records that are often asynchronous with the original dates of publication. Thus a December 2000 article may appear in the WOS in January 2001. And due to timing variations in



Potentially missed citations
9 nodes have citations that may potentially refer to other nodes.

12 1965 ASLIB PROCEEDINGS 17(6):184-196
MARTYN J
AN EXAMINATION OF CITATION INDEXES

MARTYN J, 1964, J DOC, V20, P212 may refer to 7 MARTYN-J-1964-V20-P236

42 1968 LIBRARY RESOURCES & TECHNICAL SERVICES 12(4):415-& ←
HUANG TS
EFFICACY OF CITATION INDEXING IN REFERENCE RETRIEVAL

BROWN SC, 1966, PHYSICS TODAY, V19, P60 may refer to 24 BROWN-SC-1966-V19-P59
BROWN SC, 1966, PHYSICS TODAY, V19, P61 may refer to 24 BROWN-SC-1966-V19-P59
BROWN SC, 1966, PHYSICS TODAY, V19, P64 may refer to 24 BROWN-SC-1966-V19-P59
GARFIELD E, 1963, AM DOC, V14, P290 may refer to 1 GARFIELD-E-1963-V14-P289
GARFIELD E, 1964, SCIENCE, V144, P651 may refer to 9 GARFIELD-E-1964-V144-P649

67 1971 PHYSICS TODAY 24(7):28-&
ZUCKERMAN H; MERTON RK
SOCIOLOGY OF REFEREEING

COLE S, 1968, AM SOCIOLOGICAL REV, V33, P412 may refer to 33 COLE-S-1968-V33-P397

102 1978 SCIENTOMETRICS 1(1):9-34
GILBERT GN
MEASURING THE GROWTH OF SCIENCE - REVIEW OF INDICATORS OF SCIENTIFIC GROWTH

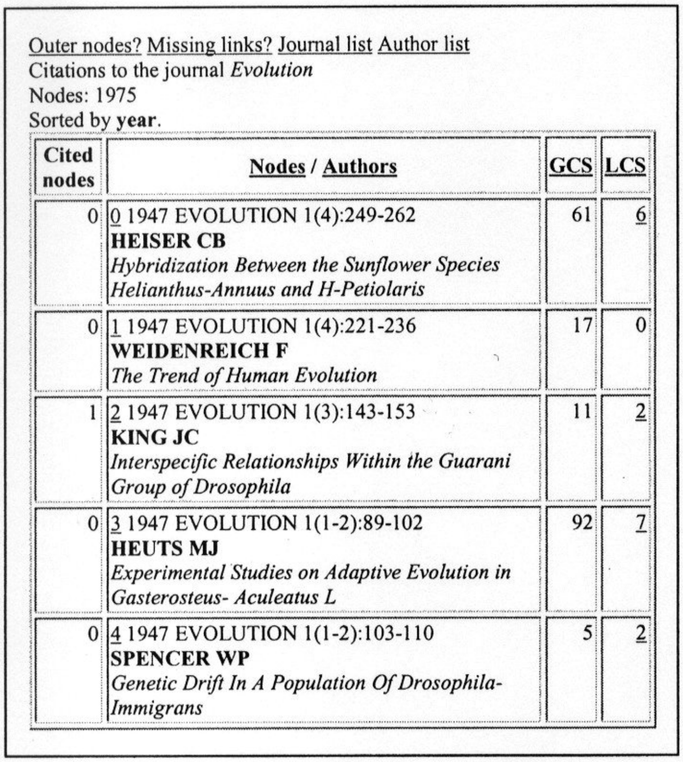
COLE S, 1968, AM SOC REV, V33, P297 may refer to 33 COLE-S-1968-V33-P397
GARFIELD E, 1963, AM DOC, V14, P290 may refer to 1 GARFIELD-E-1963-V14-P289

131 1984 NAUCHNO-TEKHNICHESKAYA INFORMATSIYA SERIYA 1-ORGANIZATSIYA I METODIKA INFORMATSIONNOI RABOTY (12):6-11
SHARABCHIEV YT
APPLICATION OF CLUSTER-ANALYSIS IN SCIENTIFIC INVESTIGATIONS

KESSLER MM, 1963, J AM DOC, V14, P99 may refer to 0 KESSLER-MM-1963-V14-P10

Figure 7: Missing links

adding back issues to ISI files, it is not possible to rely on the WOS, at present, to produce an absolute chronological arrangement.



Outer nodes? Missing links? Journal list Author list
Citations to the journal *Evolution*
Nodes: 1975
Sorted by **year**.

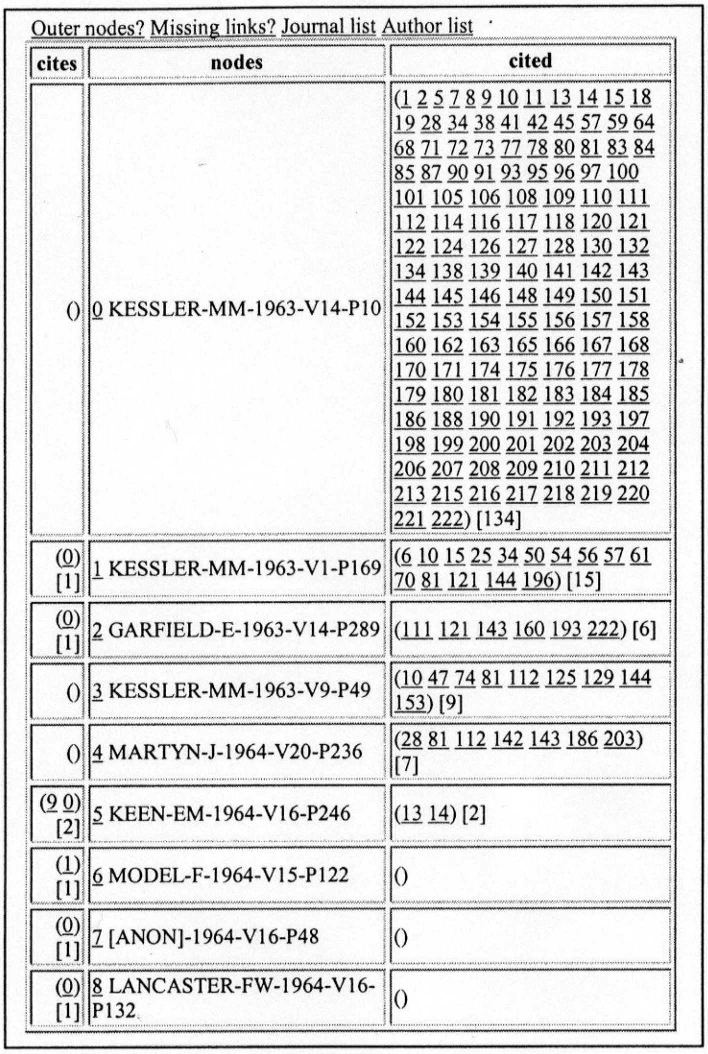
Cited nodes	Nodes / Authors	GCS	LCS
0	0 1947 EVOLUTION 1(4):249-262 **HEISER CB** *Hybridization Between the Sunflower Species Helianthus-Annuus and H-Petiolaris*	61	6
0	1 1947 EVOLUTION 1(4):221-236 **WEIDENREICH F** *The Trend of Human Evolution*	17	0
1	2 1947 EVOLUTION 1(3):143-153 **KING JC** *Interspecific Relationships Within the Guarani Group of Drosophila*	11	2
0	3 1947 EVOLUTION 1(1-2):89-102 **HEUTS MJ** *Experimental Studies on Adaptive Evolution in Gasterosteus- Aculeatus L*	92	7
0	4 1947 EVOLUTION 1(1-2):103-110 **SPENCER WP** *Genetic Drift In A Population Of Drosophila-Immigrans*	5	2

Figure 8: Chronological file of papers published in *Evolution*

While the *HistCite* program will produce a nearly perfect sort for a single journal's records, it is not possible to create an absolutely reliable chronological file for a collection covering articles from many journals since volumes are not correlated with exact publication dates. However, the monthly or weekly dates are included in the original *WOS* records. In fast moving fields it might even be better to use the date the manuscript was received to more accurately portray the sequence of rapid developments.

Citation Matrix

In order to help the user better visualize the inter-relations between the thousands of cited papers in the network, the software creates a citation matrix which displays the nodal number for citing and cited works (see Figure 9). This matrix can then become the input for the creation of citation maps of various kinds. In the first public demonstrations of this system (Garfield, 2001), we showed some maps created by Howard White of Drexel University using his co-cited author software as well as critical path software. However, our system includes a mapping procedure that produces a pure chronological listing of the 50 core papers selected by the user.



Outer nodes? Missing links? Journal list Author list

cites	nodes	cited
()	0 KESSLER-MM-1963-V14-P10	(1 2 5 7 8 9 10 11 13 14 15 18 19 28 34 38 41 42 45 57 59 64 68 71 72 73 77 78 80 81 83 84 85 87 90 91 93 95 96 97 100 101 105 106 108 109 110 111 112 114 116 117 118 120 121 122 124 126 127 128 130 132 134 138 139 140 141 142 143 144 145 146 148 149 150 151 152 153 154 155 156 157 158 160 162 163 165 166 167 168 170 171 174 175 176 177 178 179 180 181 182 183 184 185 186 188 190 191 192 193 197 198 199 200 201 202 203 204 206 207 208 209 210 211 212 213 215 216 217 218 219 220 221 222) [134]
(0) [1]	1 KESSLER-MM-1963-V1-P169	(6 10 15 25 34 50 54 56 57 61 70 81 121 144 196) [15]
(0) [1]	2 GARFIELD-E-1963-V14-P289	(111 121 143 160 193 222) [6]
()	3 KESSLER-MM-1963-V9-P49	(10 47 74 81 112 125 129 144 153) [9]
()	4 MARTYN-J-1964-V20-P236	(28 81 112 142 143 186 203) [7]
(9 0) [2]	5 KEEN-EM-1964-V16-P246	(13 14) [2]
(1) [1]	6 MODEL-F-1964-V15-P122	()
(0) [1]	7 [ANON]-1964-V16-P48	()
(0) [1]	8 LANCASTER-FW-1964-V16-P132	()

Figure 9: Citation Matrix

Historiograph for Bibliographic Coupling

Figure 10 is a manually created map on bibliographic coupling based on the data compiled by the program.

In the oral presentation, most of the figures will be demonstrated dynamically.

The subject of gene flow, is of considerable interest to one of us (Pudovkin). Instead of a cited reference search, we conducted a typical search in *WOS* using the simple title word search on "gene flow." 620 papers were published on this topic between 1974 and 2001.

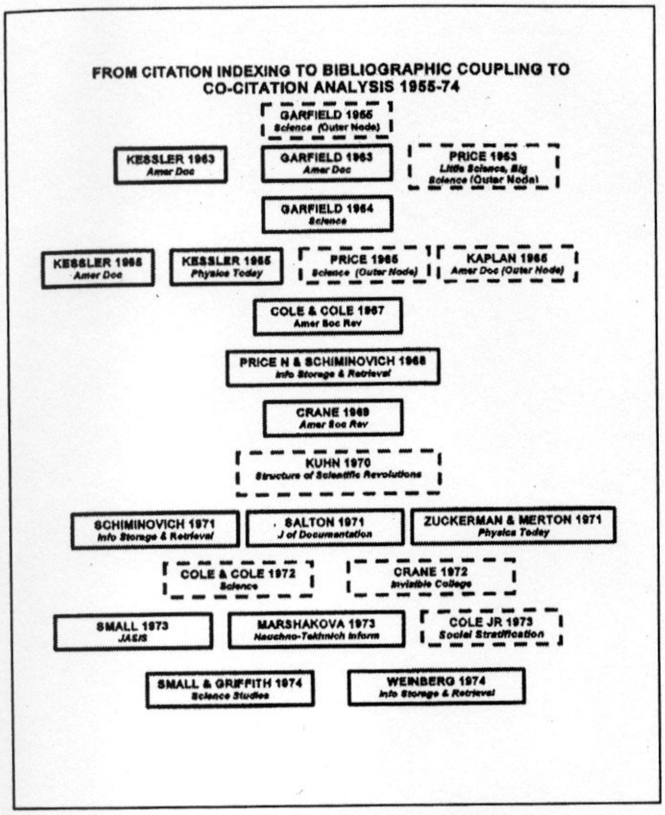
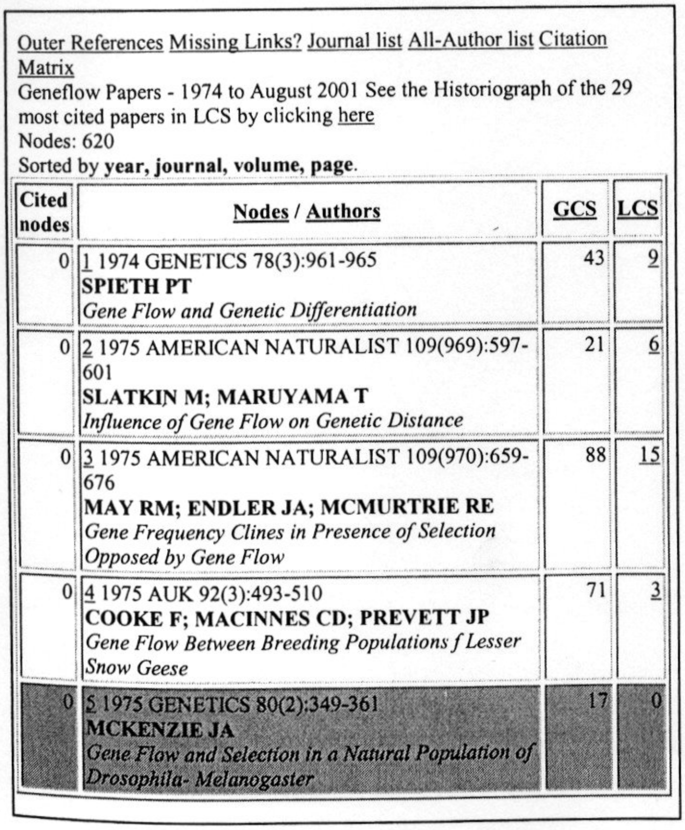

FROM CITATION INDEXING TO BIBLIOGRAPHIC COUPLING TO CO-CITATION ANALYSIS 1955-74
GARFIELD 1955 Science (Outer Node)
KESSLER 1963 Amer Doc
GARFIELD 1963 Amer Doc
PRICE 1963 Little Science, Big Science (Outer Node)
GARFIELD 1964 Science
KESSLER 1965 Amer Doc
KESSLER 1965 Physics Today
PRICE 1965 Science (Outer Node)
KAPLAN 1965 Amer Doc (Outer Node)
COLE & COLE 1967 Amer Soc Rev
PRICE N & SCHIMINOVICH 1968 Info Storage & Retrieval
CRANE 1969 Amer Soc Rev
KUHN 1970 Structure of Scientific Revolutions
SCHIMINOVICH 1971 Info Storage & Retrieval
SALTON 1971 J of Documentation
ZUCKERMAN & MERTON 1971 Physics Today
COLE & COLE 1972 Science
CRANE 1972 Invisible College
SMALL 1973 JASIS
MARSHAKOVA 1973 Nauchno-Tekhnich Inform
COLE JR 1973 Social Stratification
SMALL & GRIFFITH 1974 Science Studies
WEINBERG 1974 Info Storage & Retrieval

Figure 10: Manually created map


Outer References Missing Links? Journal list All-Author list Citation Matrix
Geneflow Papers - 1974 to August 2001 See the Historiograph of the 29 most cited papers in LCS by clicking here
Nodes: 620
Sorted by **year, journal, volume, page**.

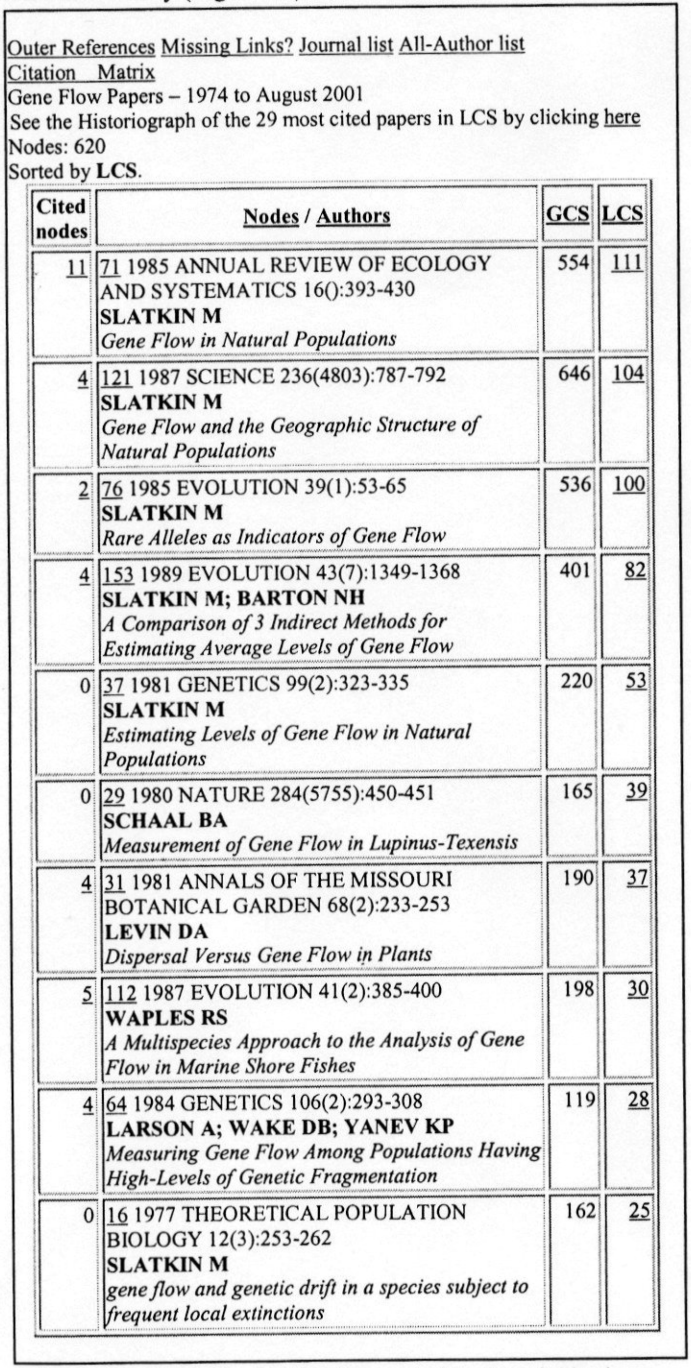
Cited nodes	Nodes / Authors	GCS	LCS
0	1 1974 GENETICS 78(3):961-965 **SPIETH PT** *Gene Flow and Genetic Differentiation*	43	9
0	2 1975 AMERICAN NATURALIST 109(969):597-601 **SLATKIN M; MARUYAMA T** *Influence of Gene Flow on Genetic Distance*	21	6
0	3 1975 AMERICAN NATURALIST 109(970):659-676 **MAY RM; ENDLER JA; MCMURTRIE RE** *Gene Frequency Clines in Presence of Selection Opposed by Gene Flow*	88	15
0	4 1975 AUK 92(3):493-510 **COOKE F; MACINNES CD; PREVETT JP** *Gene Flow Between Breeding Populations f Lesser Snow Geese*	71	3
0	5 1975 GENETICS 80(2):349-361 **MCKENZIE JA** *Gene Flow and Selection in a Natural Population of Drosophila- Melanogaster*	17	0

Figure 11: Chronological table of papers on gene flow from 1974 to 2001

Figure 11 shows a portion of the full chronological collection.

These were used to create an historiograph of the field (Figure 12). Only a portion of the 29 papers cited 10 or more times is shown.

The genealogical graphical presentation is seen in Figure 13. Note that each rectangular node is hotlinked to a full source entry (Figure 14).


Outer References Missing Links? Journal list All-Author list
Citation Matrix
Gene Flow Papers – 1974 to August 2001
See the Historiograph of the 29 most cited papers in LCS by clicking here
Nodes: 620
Sorted by **LCS**.

Cited nodes	Nodes / Authors	GCS	LCS
11	71 1985 ANNUAL REVIEW OF ECOLOGY AND SYSTEMATICS 16():393-430 **SLATKIN M** *Gene Flow in Natural Populations*	554	111
4	121 1987 SCIENCE 236(4803):787-792 **SLATKIN M** *Gene Flow and the Geographic Structure of Natural Populations*	646	104
2	76 1985 EVOLUTION 39(1):53-65 **SLATKIN M** *Rare Alleles as Indicators of Gene Flow*	536	100
4	153 1989 EVOLUTION 43(7):1349-1368 **SLATKIN M; BARTON NH** *A Comparison of 3 Indirect Methods for Estimating Average Levels of Gene Flow*	401	82
0	37 1981 GENETICS 99(2):323-335 **SLATKIN M** *Estimating Levels of Gene Flow in Natural Populations*	220	53
0	29 1980 NATURE 284(5755):450-451 **SCHAAL BA** *Measurement of Gene Flow in Lupinus-Texensis*	165	39
4	31 1981 ANNALS OF THE MISSOURI BOTANICAL GARDEN 68(2):233-253 **LEVIN DA** *Dispersal Versus Gene Flow in Plants*	190	37
5	112 1987 EVOLUTION 41(2):385-400 **WAPLES RS** *A Multispecies Approach to the Analysis of Gene Flow in Marine Shore Fishes*	198	30
4	64 1984 GENETICS 106(2):293-308 **LARSON A; WAKE DB; YANEV KP** *Measuring Gene Flow Among Populations Having High-Levels of Genetic Fragmentation*	119	28
0	16 1977 THEORETICAL POPULATION BIOLOGY 12(3):253-262 **SLATKIN M** *gene flow and genetic drift in a species subject to frequent local extinctions*	162	25

Figure 12: Gene flow collection sorted by Local Citation Score (LCS)

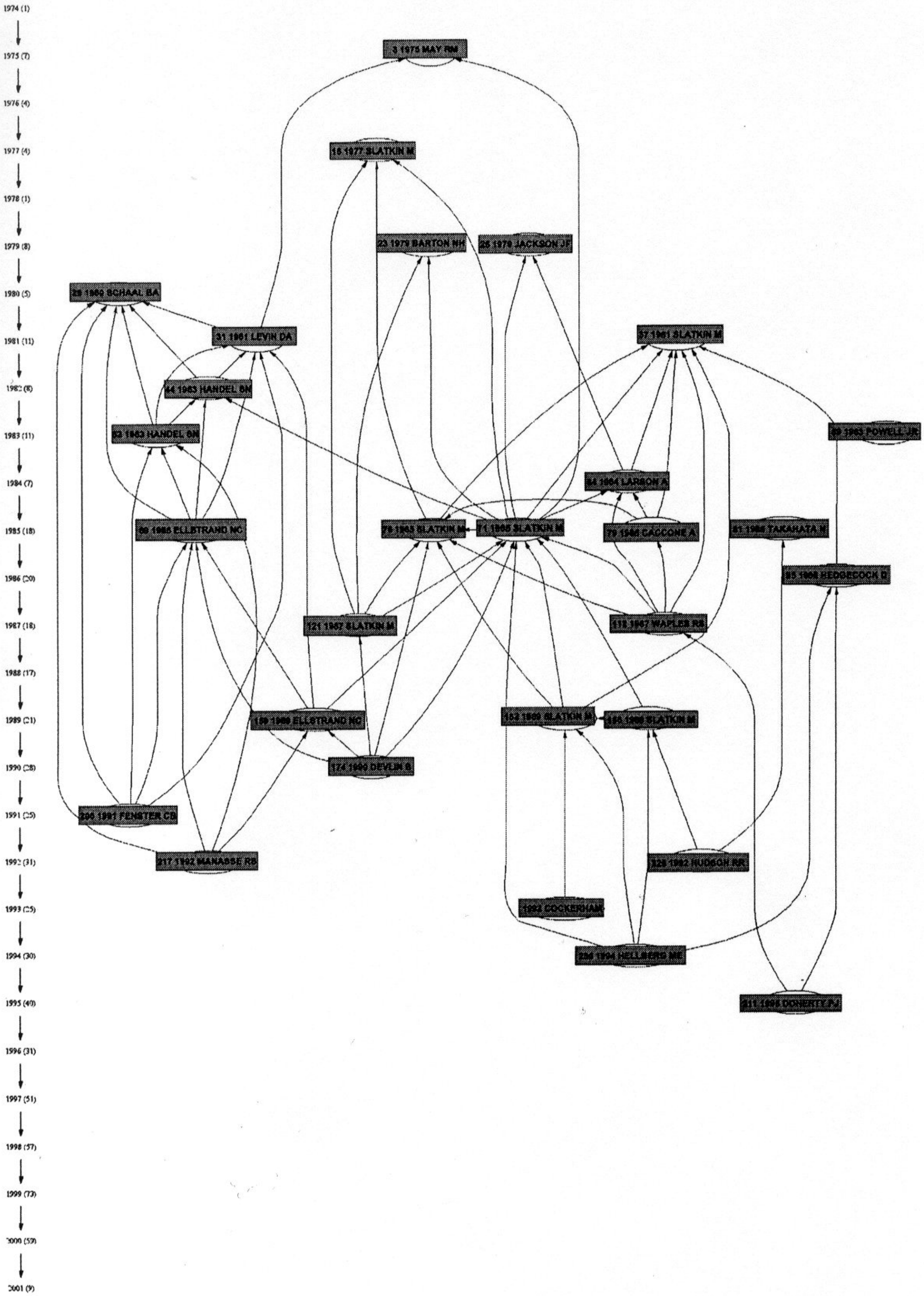
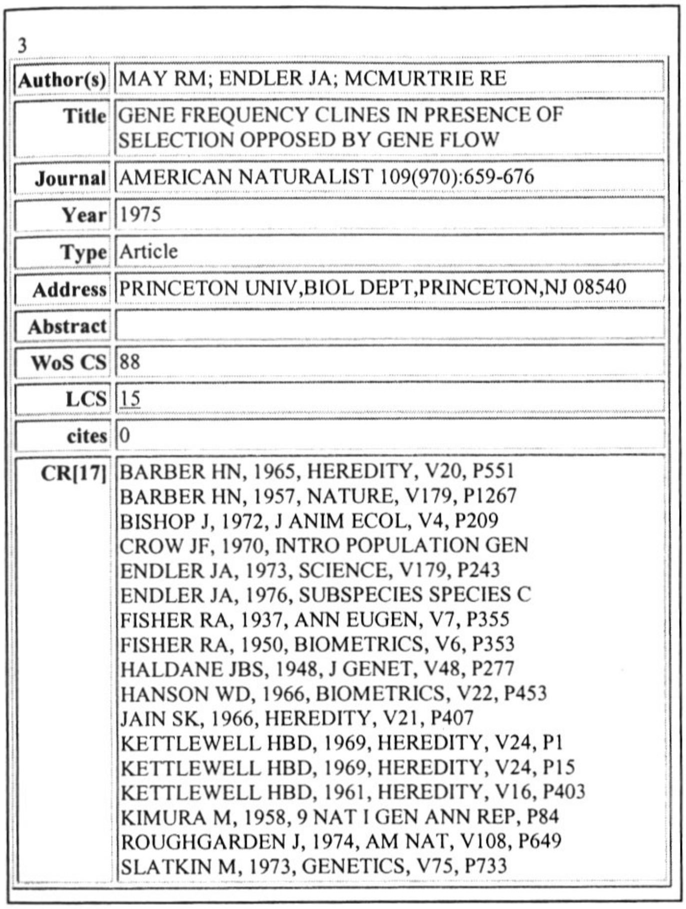


Figure 13: Computer generated historiography of "gene flow" most-cited papers



3

Author(s)	MAY RM; ENDLER JA; MCMURTRIE RE
Title	GENE FREQUENCY CLINES IN PRESENCE OF SELECTION OPPOSED BY GENE FLOW
Journal	AMERICAN NATURALIST 109(970):659-676
Year	1975
Type	Article
Address	PRINCETON UNIV,BIOL DEPT,PRINCETON,NJ 08540
Abstract	
WoS CS	88
LCS	15
cites	0
CR[17]	BARBER HN, 1965, HEREDITY, V20, P551 BARBER HN, 1957, NATURE, V179, P1267 BISHOP J, 1972, J ANIM ECOL, V4, P209 CROW JF, 1970, INTRO POPULATION GEN ENDLER JA, 1973, SCIENCE, V179, P243 ENDLER JA, 1976, SUBSPECIES SPECIES C FISHER RA, 1937, ANN EUGEN, V7, P355 FISHER RA, 1950, BIOMETRICS, V6, P353 HALDANE JBS, 1948, J GENET, V48, P277 HANSON WD, 1966, BIOMETRICS, V22, P453 JAIN SK, 1966, HEREDITY, V21, P407 KETTLEWELL HBD, 1969, HEREDITY, V24, P1 KETTLEWELL HBD, 1969, HEREDITY, V24, P15 KETTLEWELL HBD, 1961, HEREDITY, V16, P403 KIMURA M, 1958, 9 NAT I GEN ANN REP, P84 ROUGHGARDEN J, 1974, AM NAT, V108, P649 SLATKIN M, 1973, GENETICS, V75, P733

Figure 14: Full source entry for node #3, paper by R. M. May, *American Naturalist*, 1975.

Node #3 at the top of Figure 13 is Richard May's 1975 paper in the *American Naturalist*.

In Figure 15 (next page), the source set of 29 papers is shown in another format. Each document is represented by a circle whose area is proportional to its citation frequency. Ultimately, both characteristics will be included in the map.

Time and space do not permit us to discuss "Why Do We Need Algorithmic Historiography?" A paper on this theme has been accepted for the forthcoming special issue of the *Journal of the American Society for Information Science and Technology* on "Visualization of Scientific Paradigms."

Conclusion

In conclusion, we have described a tool which permits the user to manage the voluminous references produced in a comprehensive search of the literature. For those who are new to the subject, the mere juxtaposition of the most-cited papers for each five- or ten-year period of the literature will help identify the key literature to be used first. For those who are knowledgeable in the field, the system will help jog the memory to recall the key works which were associated with the development of the field. While the relevance of citing works may be apparent, the collective bibliographic coupling and co-citation of papers in and outside the basic bibliography should provide a comprehensive structure for completing a synoptic history of the topic.

Due to space restraints, only a portion of the tables shown during the oral presentation are shown here. For the complete slides, and for more readable versions of the graphs, please go to http://www.garfield.library.upenn.edu/papers/ASIS2002presentation.html.

References

Buter, R.K. and Noyons, E.C.M. (2001). Improving the functionality of interactive bibliometric science maps. Scientometrics 51, 55-68.

Cawkell, A.E. (October 30, 1989). Acoustic journals and acoustic research articles. Current Contents No. 44, 4-15. Reprinted in Garfield, E. (1977) Essays of an Information Scientist, (Vol. 12, pp. 4-15). Philadelphia: ISI Press. Available: http://www.garfield.library.upenn.edu/essays/v12p301y1989.pdf

Cawkell, A.E. (2000). Visualizing Citation Connections. In B. Cronin & H.Barsky Atkins (Eds), The Web of Knowledge: A Festschrift in Honor of Eugene Garfield (ASIS Monographic Series) (pp. 177-194). Medford, NJ: Information Today, Inc.

Garfield, E., Sher I.H., & Torpie R.J. (December 1964). The Use of Citation Data in Writing the History of Science. Report of research for Air Force Office of Scientific Research under contract AF49(638)-1256. Philadelphia: The Institute for Scientific Information, December 1964. Available: http://www.garfield.library.upenn.edu/papers/useofcitdatawritinghistofsci.pdf

Garfield, E. (April 14, 1971) Citation indexing, historio-bibliography and the sociology of science biography. Current Contents, No. 15, pages M25+. Reprinted from: K.E. Davis & W.D. Sweeney (Eds) Proceedings of the Third International Congress of Medical Librarianship 5-9 May 1969. (pp. 187-204). Amsterdam: Excerpta Medica. Reprinted in Garfield E. (1977). Essays of an Information Scientist (Vol 1, pp.158-174). Philadelphia: ISI Press (1977). Available: http://www.garfield.library.upenn.edu/essays/V1p158y1962-73.pdf

Garfield, E. (June 8, 1992). Contract Research Services at ISI—Citation Analysis for governmental, Industrial, and Academic Clients, Current Contents No. 23, .5-13. Reprinted in Garfield E. (1993). Essays of an Information Scientist (Vol. 15, pp. 75-83). Philadelphia: ISI Press (1993). Available: http://garfield.library.upenn.edu/essays/v15p075y1992-93.pdf

Garfield, E. (September 19, 2001) From computational linguistics to algorithmic historiography, paper presented at the Symposium in Honor of Casimir Borkowski at the University of Pittsburgh School of Information Sciences. Available: http://garfield.library.upenn.edu/papers/pittsburgh92001.pdf

Lawrence S, Giles, CL., & Bollacker K. (1999). Digital libraries and autonomous Citation Indexing, Computer 32, 67-71. Available: http://www.neci.nec.com/~lawrence/papers/aci-computer99/

Small, H. and Garfield, E. (1985). The geography of science: disciplinary and national mappings, Journal of Information Science, 11, 147-159 (1985). Reprinted in Garfield E. (October 27, 1986) Current Contents No. 43, 3-14. Reprinted in Garfield E. (1988). Essays of an Information Scientist. (Vol. 9, pp. 324-335). Philadelphia: ISI Press. Available: http://www.garfield.library.upenn.edu/essays/v9p324y1986.pdf

Small, H. (1994). A Sci-Map case study: Building a map of AIDS Research. Scientometrics, 30, :229-241 (1994)

Small, H. , Sweeney, E., and Greenlee, E. (1985). Clustering the Science Citation Index using co-citations. 2. mapping science. Scientometrics, 8, 321-34

White, H.D. and McCain K.W. (1997). Visualization of literatures. Annual Review of Information Science and Technology, 32, 99-168.

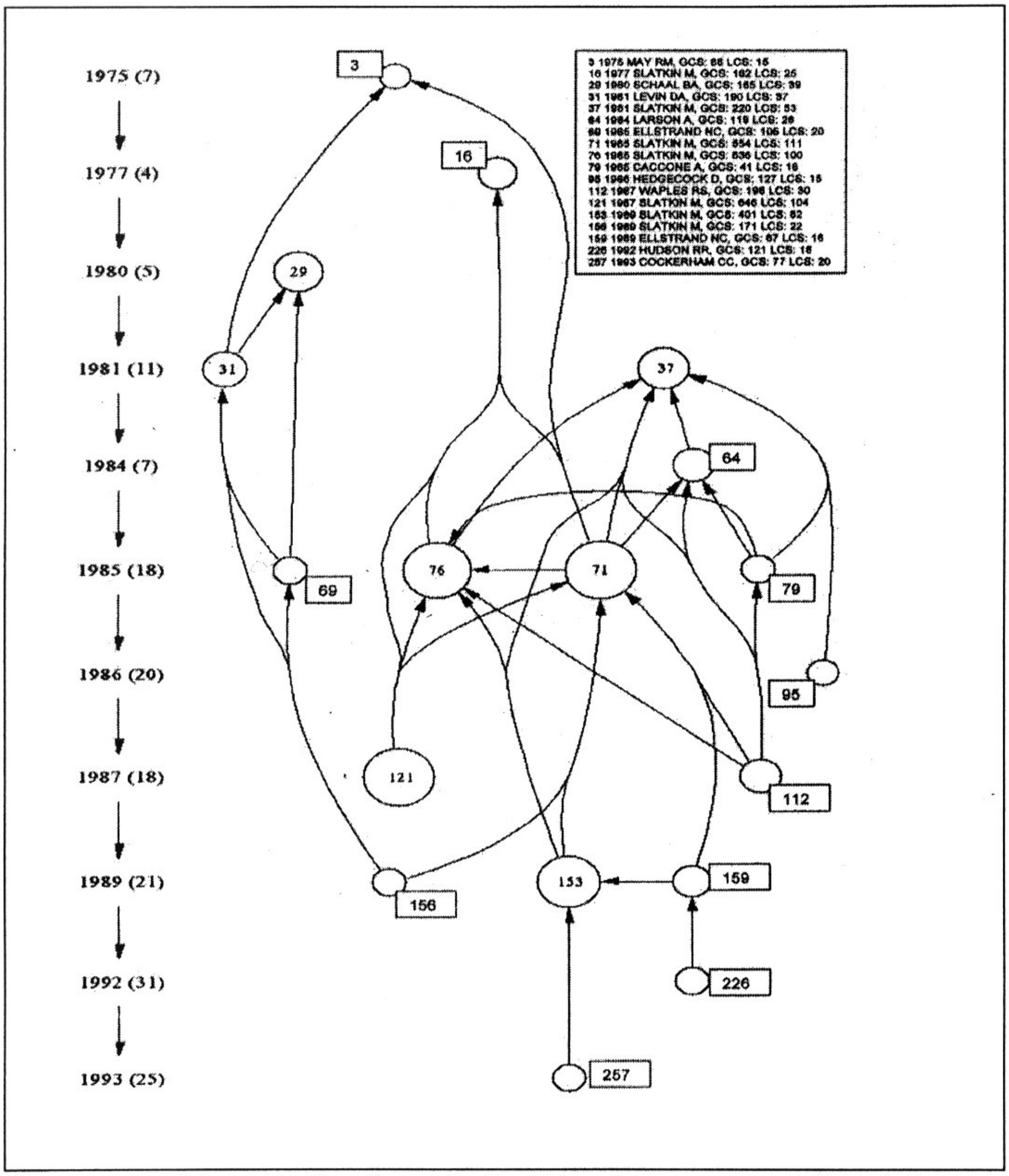

1975 (7)
1977 (4)
1980 (5)
1981 (11)
1984 (7)
1985 (18)
1986 (20)
1987 (18)
1989 (21)
1992 (31)
1993 (25)
3
16
29
31
37
64
69
76
71
79
95
121
112
156
153
159
226
257
3 1975 MAY RM, GCS: 86 LCS: 15
16 1977 SLATKIN M, GCS: 182 LCS: 25
29 1980 SCHAAL BA, GCS: 165 LCS: 39
31 1981 LEVIN DA, GCS: 190 LCS: 37
37 1981 SLATKIN M, GCS: 220 LCS: 53
64 1984 LARSON A, GCS: 119 LCS: 28
69 1985 ELLSTRAND NC, GCS: 106 LCS: 20
71 1985 SLATKIN M, GCS: 854 LCS: 111
76 1985 SLATKIN M, GCS: 836 LCS: 100
79 1985 CACCONE A, GCS: 41 LCS: 16
95 1986 HEDGECOCK D, GCS: 127 LCS: 15
112 1987 WAPLES RS, GCS: 196 LCS: 30
121 1987 SLATKIN M, GCS: 646 LCS: 104
153 1989 SLATKIN M, GCS: 401 LCS: 82
156 1989 SLATKIN M, GCS: 171 LCS: 22
159 1989 ELLSTRAND NC, GCS: 67 LCS: 16
226 1992 HUDSON RR, GCS: 121 LCS: 18
257 1993 COCKERHAM CC, GCS: 77 LCS: 20

Figure 15: Gene flow collection with each paper represented by a circle proportional to citation frequency