

Crossmaps: visualization of overlapping relationships in collections of journal papers

Steven A. Morris*, Gary G. Yen*

*Electrical and Computer Engineering, Oklahoma State University, 202 Engineering So., Stillwater, OK 74078, E-mail: samorri@okstate.edu

A crossmapping technique is introduced for visualizing multiple and overlapping relations among entity types in collections of journal articles. Groups of entities from two entity-types are crossplotted to show correspondence of relations. For example, author collaboration groups are plotted on the x-axis against groups of papers (research fronts) on the y-axis. At the intersection of each pair of author group/research front pairs a circular symbol is plotted whose size is proportional to the number of times that authors in the group appear as authors in papers in the research front. Entity groups are found by agglomerative hierarchical clustering using conventional similarity measures. Crossmaps comprise a simple technique that is particularly suited to showing overlap in relations among entity groups. Particularly useful crossmaps are: research fronts against base reference clusters, research fronts against author collaboration groups, and research fronts against term co-occurrence clusters. When exploring the knowledge domain of a collection of journal papers, it is useful to have several crossmaps of different entity pairs, complemented by research front timelines and base reference cluster timelines.

Introduction. Collections of journal papers related to a scientific field are a useful source of information when mapping a knowledge domain (1). The structure within the knowledge domain is manifested in the collection of papers as groups of related entities, such as groups of papers that represent subtopics, groups of references that represent base knowledge, groups of paper authors that represent collaboration teams, groups of reference authors that represent experts, groups of journals that represent subtopic libraries, and groups of terms that represent specialized vocabularies within the knowledge domain. Exploration and visualization of these groups and the complex relations among them provides information that can be used to gain a broad and detailed understanding of the underlying knowledge domain.

Inherently, entity groups within collections of journal papers exhibit considerable “core and scatter” in group membership (2), with each group usually possessing a small core group of strongly related member entities, and a much larger group of weakly related “scatter” members. Furthermore, there is considerable overlap in membership of entities in groups. For a thorough understanding of the structure of a knowledge domain, it is useful to visualize and understand the extent of overlap among groups in a collection of journal papers.

This paper introduces a simple technique for visualizing the relations among collections of entity groups. The technique, which uses a crossmap format to show the magnitude of correspondences between all pairs of groups drawn from two differing entity-types, allows visualization of relations between groups and additionally permits visualization of overlap in group membership. Using this technique it is possible to visualize and understand the set of complex relations among the different groups that are manifested in a knowledge domain. For example, given a collection of research fronts, i.e., groups of papers reporting on the same sub-topic, for each research front it is possible to identify the groups of important references, contributing author collaboration teams, groups of experts (important reference authors), and key journals.

This paper is organized as follows. The reader is introduced to a simple entity-relationship model of collections of journal papers. This is followed by a discussion of important entities used for mapping knowledge

domains and by a discussion of co-occurrence relations that are used to cluster entities into groups. A detailed discussion of important entity groups appears, including an explanation of core and scatter and overlap of group membership. After this the use of correspondence metrics between entity groups is discussed, followed by a detailed discussion of the proposed crossmapping technique. Finally, an example is presented, showing crossmaps produced from a collection of journal papers related to the subject of complex networks.

Entity model of journal paper collections. Using an entity-relationship model (3), collections of journal papers may be considered to be a collection of entities of differing entity-types. Examples of entity-types for journal paper collections include papers, paper authors, references, and paper journals. Borner, Chen, and Boyack (1) describe these entities as “units of analysis” and list the entity-types most commonly used for mapping knowledge domains and also show applications of the analysis of each entity-type. This paper will expand upon analysis of entities for knowledge mapping to include analysis of relations between different types of entities, thus extending the understanding of complex knowledge domains.

Within the collection of journal papers, entities are associated with each other. For example, each paper in the collection is associated with the authors who wrote it, the references it cites, the journal in which it was published and the terms that appear in it. As presented here, these associations are always between pairs of entities of differing entity-type. Entities of the same entity type are never directly associated. For example, papers and references are considered distinct entity-types, even though references often correspond to actual papers. This separation into two distinct entity-types is necessary because papers and references represent differing concepts. A paper represents a research report while a reference represents a symbol of knowledge (4). Similar considerations require designation of paper journals, reference journals, paper authors and reference authors as separate entity-types. Figure 1 shows an ontology diagram of the entities within a collection of journal papers and the types of associations among those entities. Figure 2 explains the symbols used in Figure 1.

Co-occurrence among entities. Co-occurrence relations among entities of the same entity-type occur when two entities of the same type are associated with an entity of a differing entity-type. For example, two authors are related when they coauthor a paper, two papers are related when both cite the same reference, or two references are related when they are cited together in the same paper. Co-occurrence relations between pairs of entities often imply some meaningful relation between those entities. For example, coauthorship of papers implies that pairs of authors are collaborators, common references between papers implies that pairs of papers deal with the same research topic, co-citation of references implies that two references are symbols of similar base knowledge.

Several of the co-occurrence relations that occur within collections of journal papers have been named and extensively studied. These relations are noted in Figure 1 and include:

- **Bibliographic coupling.** Relation of pairs of papers by common references, implying a common research topic between the papers (5).
- **Co-citation.** Relation of pairs of references by their co-occurrence in papers, implying that those two references are symbols of similar base knowledge (6).

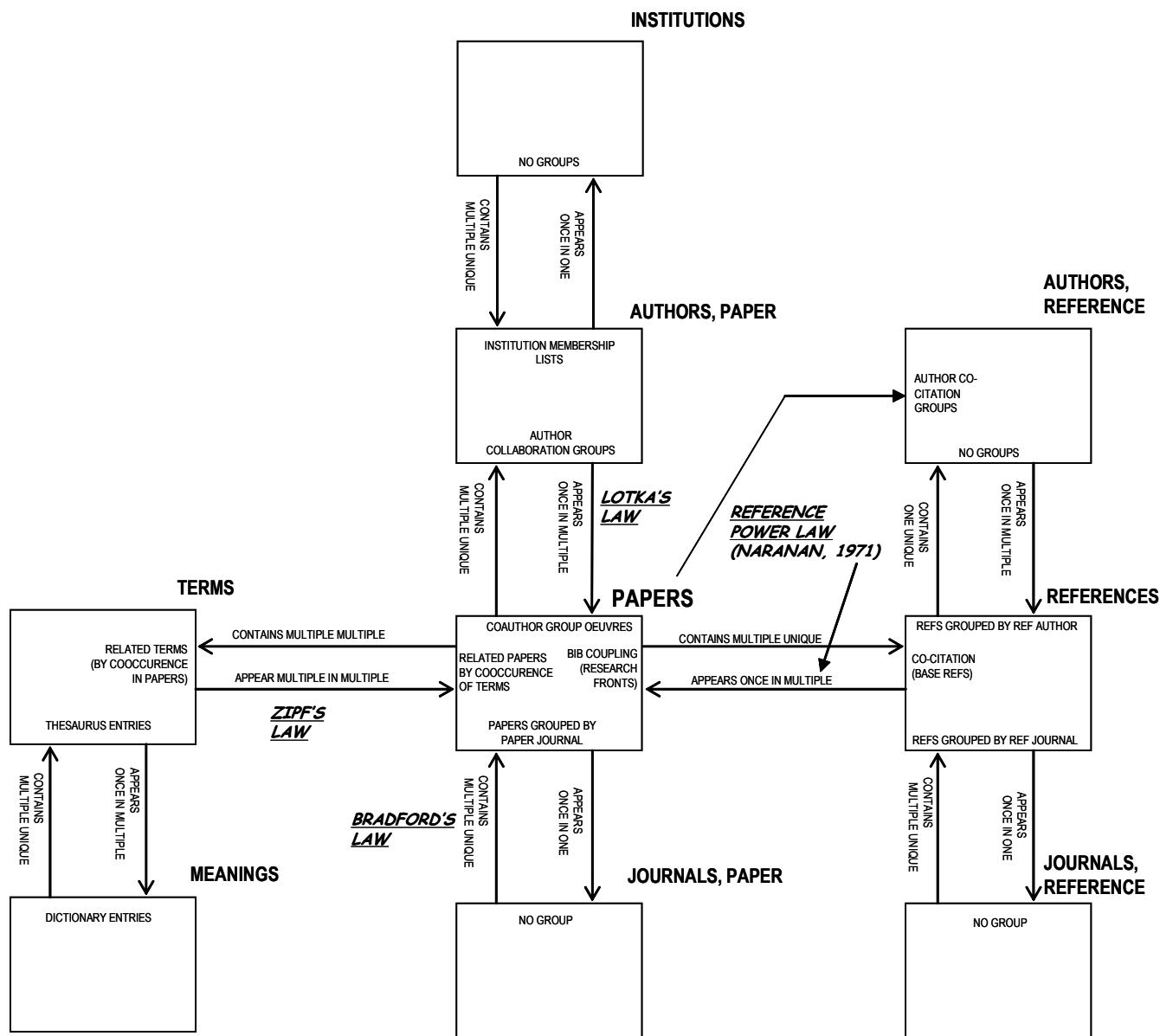


Fig. 1. Ontology diagram of the entities within a collection of journal papers, their direct relations to each other, co-occurrence groups, and core and scatter relations among those entities.

- **Author co-citation.** Relation of pairs of reference authors by their co-occurrence in papers, implying that the two authors are symbols of the same base knowledge (7).
- **Co-authorship.** Relation of pairs of paper authors by co-authorship of papers, implying that the two authors are members of the same collaboration team (8).

Many other co-occurrence relations are possible, as noted in Figure 1. For example, pairs of papers related by common terms may imply a common research topic, or pairs of journals that contain papers that cite common reference authors may imply that those two journals publish papers that have a common research topic.

Entity Groups. Using similarity metrics derived from co-occurrence counts between pairs of entities, and applying clustering techniques, groups of entities possessing commonalities can be identified in the collection of journal papers. Examples of commonly studied groups of entities are noted in Figure 1 and include (2):

- **Research fronts.** Groups of papers that share a common research topic (9). Derived from co-occurrence of references in papers, these groups can be considered as representing Kuhnian puzzles within a scientific field (10, 11).
- **Base reference groups.** Groups of references that serve as symbols of similar base knowledge (12). Derived from co-occurrence of references in papers, these groups can be considered as representing Kuhnian exemplars or paradigms.
- **Reference author groups.** Groups of reference authors that serve as symbols of similar base knowledge (13). Derived from co-occurrence of reference authors in papers. Similar to base reference groups, these groups can also be considered as representing Kuhnian exemplars and paradigms, but on a more abstract scale. Reference author groups can also be considered as groups of experts (14).

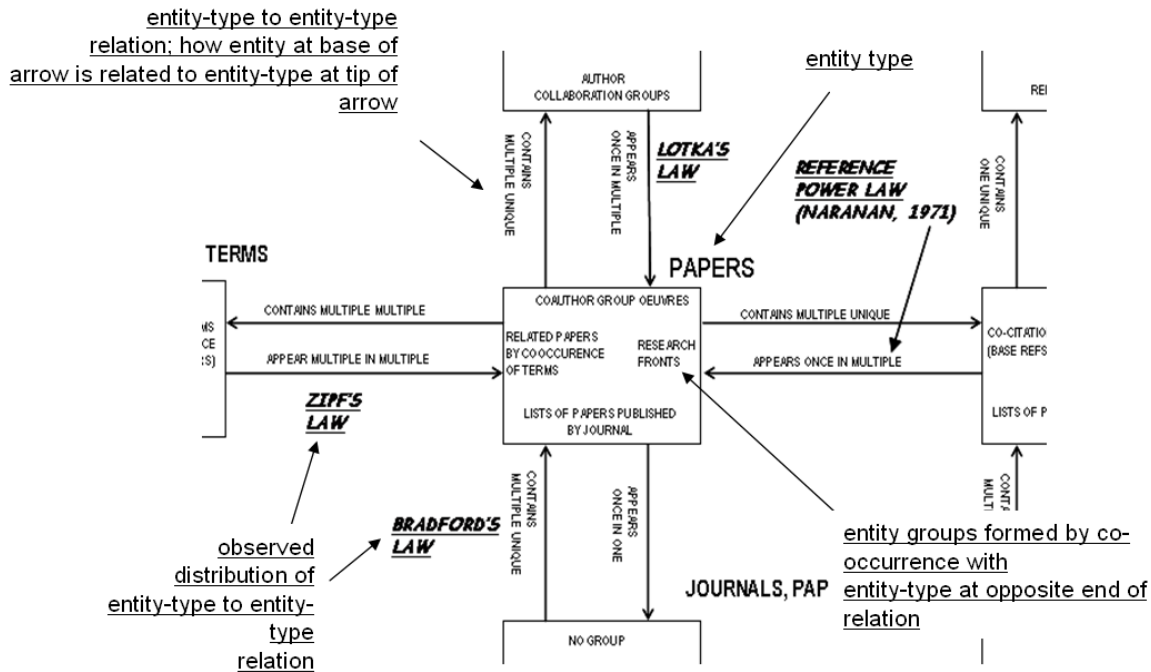


Fig. 2. Key explaining the notation in the ontology diagram of Figure 1.

- **Collaboration teams.** Groups of paper authors that work together. Derived from co-authorship of papers by paper authors, these groups can be considered as representing “invisible colleges” within a field (15, 16).
- **Vocabularies** Groups of keyword terms. Derived from co-occurrence of terms in papers, these groups can be considered to represent specialized vocabularies within a research field (17).

“Core and scatter” and overlap of group membership. Entity groups within collections of papers exhibit “core and scatter.” Groups tend to possess a small set of core members that are strongly related to each other and a large number of scatter members that are weakly related (2). Furthermore, weakly related member entities are ambiguously related to many groups simultaneously. There is extensive overlap in group membership, leading to great difficulties when visualizing the knowledge domain represented by a collection of journal papers.

The core and scatter relations among entities in collections of journal papers manifest themselves as power-law distributions of entity frequency. Some of these relations, noted on Figure 1, have been extensively researched. Example power-law relations are:

- Lotka’s Law for author frequency (18).
- Bradford’s Law for paper journal frequency (19).
- Zipf’s Law for frequency of terms (20).
- Reference power law for frequency of references (21).

Standard clustering techniques, such as hierarchical agglomerative clustering, and standard visualization techniques, such as multidimensional scaling, do not effectively reveal the overlap of entity group membership in collections of journal papers. The crossmapping technique proposed here is designed to reveal this overlap in a field’s knowledge structure by showing overlap in correspondence among groups taken from differing entity-types.

Correspondence between groups of entities from differing entity-types. Define a correspondence metric to measure the relation between a pair of entity groups drawn from differing entity-types. As an example, a possible correspondence metric between a research front (a group of

papers) and a base reference group is the percentage of references in the base reference group that are cited by papers in the research front. Given two collections of groups, each collection drawn from a different entity-type, it is possible to build a matrix of correspondences that exist from each group of the first entity-type to each group of the second entity-types.

Entity groups overlap in correspondence between groups from different entity-types, e.g., a base reference cluster may have correspondence to several research fronts, or an author collaboration group may have correspondence to many term co-occurrence clusters. Knowledge of the correspondence between groups drawn from different entity-types is helpful for mapping the knowledge domain associated with the collection of journal papers. Furthermore, visualization of overlapping group-to-group correspondence helps sort out complex relations among research topics, base reference groups, and research teams. In collections of journal papers, we propose several correspondence relations between groups of different entity-types that are useful for knowledge mapping:

- **Relation of research fronts to base reference groups.** This shows what base knowledge supports specific research topics.
- **Relation of research fronts to author collaboration groups.** This shows what research teams work on specific research topics.
- **Relation of research fronts to term co-occurrence groups.** This shows what concepts are associated with specific research topics and can be helpful for labeling research fronts.
- **Relation of research fronts to paper journal groups.** This shows the core journals that publish papers pertinent to specific research topics.

The crossmapping technique presented here is used to visualize and explore relations between groups of entities. The technique is especially suitable to the visualization of overlap in such relations, and as such, allows the investigation of a knowledge domain through various manifestations: research fronts, base reference groups, invisible colleges, technical vocabularies and core journals.

Description of crossmap visualization. The crossmapping technique presented here visualizes the matrix of correspondence magnitudes between groups from two different entity-types. Assuming, for example, groups of papers (entity-type 1), and groups of references (entity-type 2), one measure of correspondence is the number of references in a group of references that appears in a group of papers. Given N_1 groups of entity-type 1 and N_2 groups of entity-type 2, a N_1 by N_2 matrix lists all of the correspondences between entity-type 1 groups and entity-type 2 groups. A crossmap is a visual representation of that correspondence matrix. The crossmap method is similar to Matrix Browser (22), used for visualizing computer networks, and DocCube (23), which uses three dimensional matrix visualizations to aid query searches of large document collections, and also GRIDL (24), an interactive system for visualizing hierarchically organized databases and library search results. The crossmap technique complements timeline visualization (11), allowing a thorough exploration of the static relations and temporal events in a collection of research fronts.

Construction of crossmaps. To start, clustering is performed on each entity-type, grouping entities according to some similarity metric. Clusters from the first entity-type are mapped as rows while clusters of the other entity-type are mapped as columns. Dendrograms are added to the crossmap to show the structure of clusters being displayed. For every group at row i from entity-type 1, and every group at column j from entity-type 2, a circle is placed at row i and column j whose size is proportional to the magnitude of the correspondence between those two groups. Group labels are placed at row and column positions to the left and bottom of the map.

Example – A collection of complex networks papers. A collection of papers about complex networks will illustrate the use of crossmap techniques. This collection was gathered from the Institute for Scientific Information (ISI) *Web of Science* product by using queries to gather papers that cite several key references in the field. All groups in this set were generated using agglomerative hierarchical clustering with Ward's method linkage on co-occurrence metrics normalized using the cosine formula (25). Summary statistics for the entities in this collection are:

- 323 papers in 86 journals. (20% of journals contain 76% of all papers)
- 11304 citations to 6167 references. (20% of references received 54% of all citations)
- 826 authorships of 455 authors. (20% of authors accounted for 52% of all authorships)

Using bibliographic coupling (5) as a co-occurrence metric, a clustering of 10 research fronts was generated. Figure 3 shows a timeline of the research fronts. Papers are shown as circles plotted by publication date in horizontal tracks whose vertical position corresponds to positions on the clustering dendrogram shown to the left. Circle size is proportional to the number of citations received, and circles are darkened for papers that have received 8 or more citations in the last 12 months. The 6 most cited papers are noted in the figure.

The research front labels were added after manually browsing titles of papers in each research front for themes. In future research, it may be possible to automate or semi-automate the labeling process by using correspondence to terms derived from term co-occurrence groups. When browsing titles, considerable overlap is found in themes. Note, for example, label 'epidemics' is used for both research front 8 and research front 10. Interestingly, research front 7, 'social networks', has many recent papers that are currently highly cited, implying an important research topic providing current base knowledge. Research fronts 10 and 8, both labeled 'epidemics' are both recent, indicating an emerging topic of research.

Figure 4 is a crossmap of research fronts to base reference clusters, which were found using the co-citation similarity metric (6). References cited less than 20 times were discarded, leaving 50 references for clustering, shown individually. Correspondence was measured by counting the number of times a reference appears in papers in a research front. The dendrogram at the top of the map shows approximately 8 base reference clusters. The central group, references 3, 9 and 16, are used by all research fronts except research fronts 4 and 3. Note the difference in

the two 'epidemics' research fronts: research front 10 uses references by authors Albert and Pastor-Satorras (references 1, 2 and 8), while research front 8 relies heavily on references by authors Moore, Barrat, and Newman (references 25, 36 and 42). It is also easy to see that research front 7 and 8 overlap in their use of references 14 through 44, but at the same time research fronts 7 and 10 overlap, using references 5 through 34.

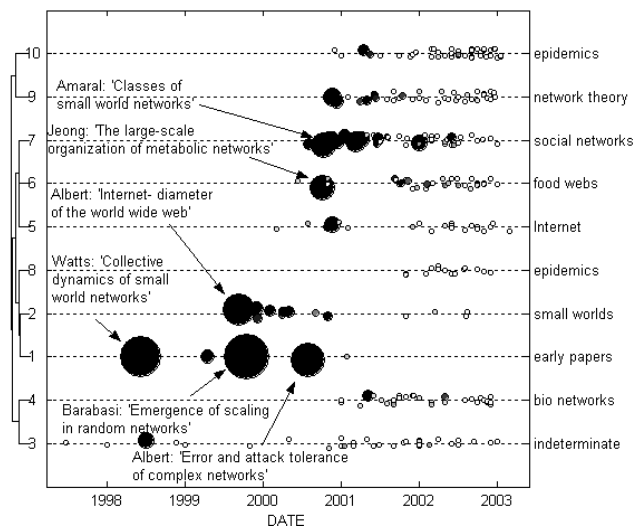


Fig. 3. Timeline of research fronts for complex networks papers. Papers are shown as circles whose size is proportional to total citations received. Darkened circles are papers that have received 8 or more citations in the last 12 months.

Figure 5 shows a crossmap of research fronts to author collaboration groups. Collaboration groups were found by using co-authorship counts as the similarity metric. Authors with less than three papers were discarded, leaving 28 authors for clustering, which are individually shown. Correspondence is measured by counting the number of times an author appears in papers in a research front. Author groups are easily discerned from the dendrogram at the top of the map. Example groups are authors 14 to 21 (Strogatz, Watts, Newman and Moore), and authors 11 to 2, (Jeong, Albert and Barabasi). Note the overlap of author group 6, 2, and 18 across research fronts 10 and 7. Additionally, Albert (author 8), whose papers were used as references from research front 10, does not appear to author any papers in that research front. Other overlapping relations are evident, particularly author groups contributing to research fronts 7 and 9.

Conclusion. The crossmapping technique shown here provides an easily understood method to explore relations in a collection of papers. In the example shown here, two types of crossmaps allow comprehension of the overlapping relations among research fronts, base reference groups, and author collaboration groups. Potential users of crossmaps are researchers exploring the literature of a scientific field to discover current research topics, base references, research teams, core journals and the relations among them. In the authors' experience the method is particularly useful for mapping a domain during the initial state-of-art review phase of new research projects when the researchers are most unfamiliar with a field's literature. The method will be useful for summarizing information about a field for presentation to subject matter experts for technology forecasting. User studies need to be conducted to validate the ease of comprehension of crossmaps and to identify the most useful pairs of entity-types for crossmapping.

The technique has so far been applied mainly to small, well-focused collections of papers (less than 1000 papers.) The principle limitation to crossmapping of large collections of papers is the restricted space available on the axes for labels. The technique may be adaptable to very large collections if interactive tools are added to expand and contract levels of the clustering hierarchy.

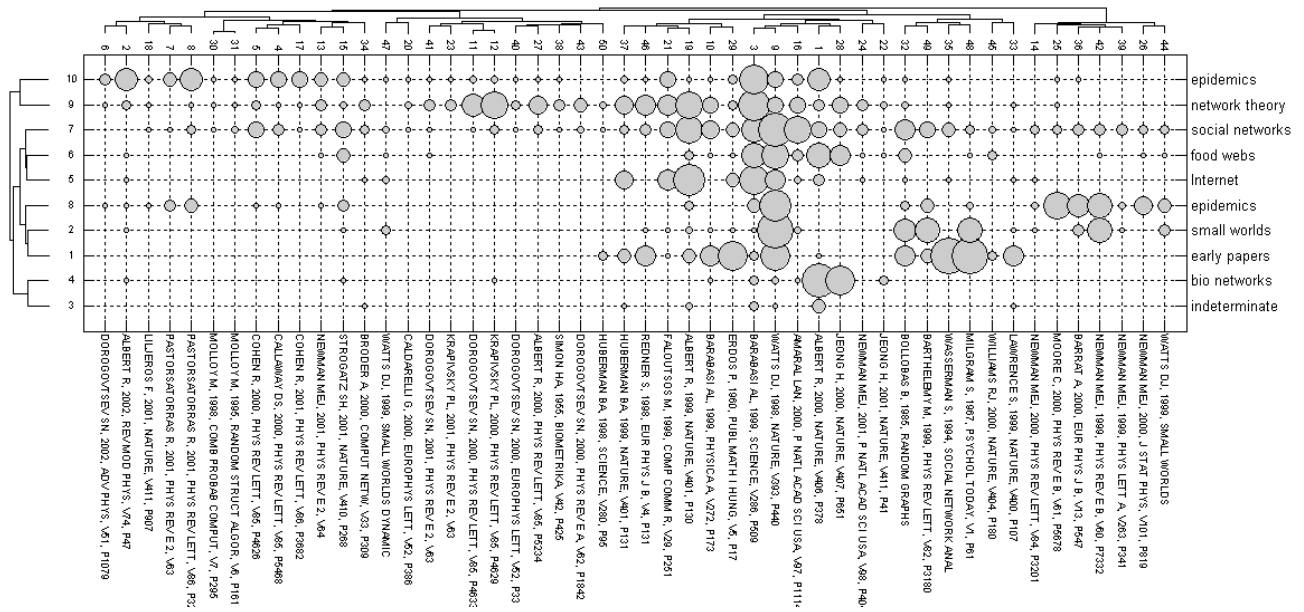


Fig. 4. Crossmap of research fronts to base reference groups for a collection of complex networks papers.

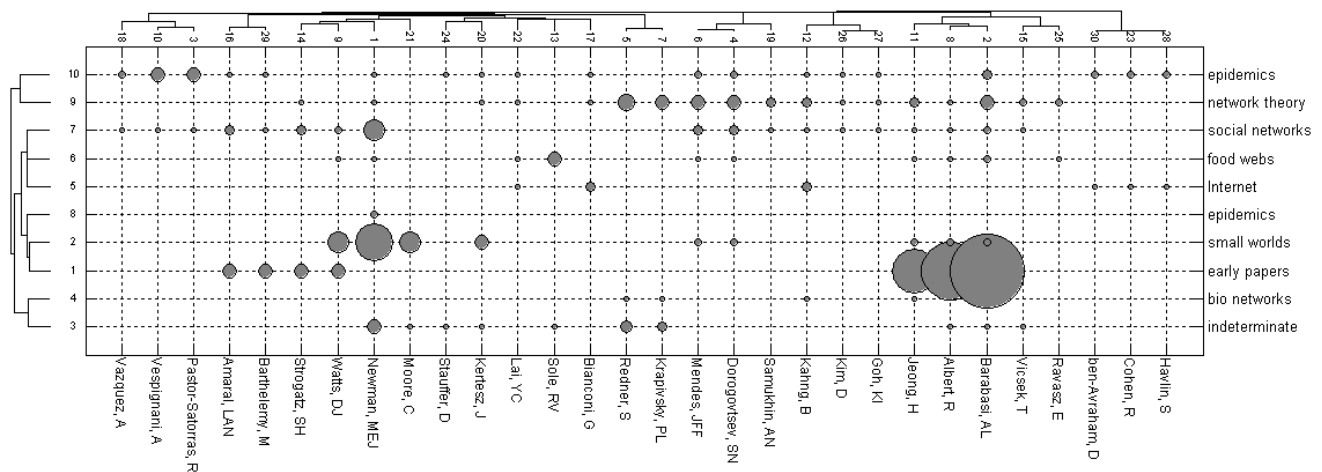


Fig. 5. Crossmap of research fronts to author collaboration groups for a collection of complex networks papers.

1. Borner, K., Chen, C. & Boyack, K. W. (2002) *Annual Review of Information Science and Technology* **37**, 179-255.
2. White, H. D. & McCain, K. W. (1989) *Annual Review of Information Science and Technology* **24**, 119-186.
3. Chen, P. P. S. (1976) *ACM Transactions on Database Systems* **1**, 9-36.
4. Small, H. (1978) *Social Studies of Science* **8**, 327-40.
5. Kessler, M. M. (1963) *American Documentation* **14**, 10-25.
6. Small, H. G. (1973) *Journal of the American Society for Information Science* **24**, 265-269.
7. White, H. D. & Griffith, B. C. (1981) *Journal of the American Society for Information Science* **32**, 163-171.
8. Subramanyam, K. (1983) *Journal of Information Science* **6**, 33-38.
9. Persson, O. (1994) *Journal of the American Society for Information Science and Technology* **45**, 31-38.
10. Kuhn, T. S. (1970) *The structure of scientific revolutions* (University of Chicago Press, Chicago.).
11. Morris, S. A., Yen, G., Wu, Z. & Asnake, B. (2003) *Journal of the American Society for Information Science and Technology* **55**, 413-422.
12. Small, H. (1997) *Scientometrics* **38**, 275-293.
13. White, H. D. & McCain, K. W. (1998) *Journal of the American Society for Information Science* **49**, 327-355.
14. Garfield, E. (1979) *Citation indexing - its theory and application in science, technology, and humanities* (Wiley, New York).
15. Kretschmer, H. (1997) *Scientometrics* **40**, 579-591.
16. Crane, D. (1972) *Invisible colleges; diffusion of knowledge in scientific communities* (University of Chicago Press, Chicago.).
17. Callon, M., Courtial, J. P. & Laville, F. (1991) *Scientometrics* **22**, 155-205.
18. Lotka, A. J. (1925) *Journal of the Washington Academy of Science* **16**, 317-323.
19. Bradford, S. C. (1938) *Engineering* **137**, 85-86.

20. Zipf, G. K. (1949) *Human behavior and the principle of least effort*. (Addison-Wesley, Reading, MA.).
21. Naranan, S. (1971) *Journal of Documentation* **27**, 83-97.
22. Ziegler, E., Kunz, C., Botsch, V. & Schneeberger, J. (2002) in *Proceedings of the IEEE Sixth International Conference on Information Visualisation* . London), pp. 361 -366.
23. Mothe, J. & Chrisment, C. (2003) *Journal of the American Society for Information Science and Technology* **54**, 650-659.
24. Shneiderman, B., Feldman, D., Rose, A. & Grau, F. G. (2000) in *Fifth ACM Conference on Digital Libraries*, San Antonio, Tx), pp. 57-66.
25. Salton, G. (1989) *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. (Addison-Wesley, Reading, Mass.).