

Mapping Institutions and Their Weak Ties in a Research Specialty: A Case Study of Cystic Fibrosis Body Composition Research¹

Liyang Yang^{2*}, Steven A. Morris^{**} and Elizabeth M. Barden^{***}

* yangly@mail.las.ac.cn

National Science Library of Chinese Academy of Science
33 Beisihuan Xilu, Zhongguancun, Beijing 100080, (P.R.China)

** steven.morris@bakerhughes.com

Baker-Hughes Inc. 2001 Rankin Road, Houston, Texas 77073 (USA)

*** embarden@comcast.net

Barden Consulting, 22 Federation Rd., Bedford, NH, 03110 (USA)

Abstract

Using a collection of papers gathered from the Web of Science on the topic of cystic fibrosis body composition research, we demonstrate analysis and visualization methods that show the collaboration structure of institutions in the specialty and the researchers that function as weak ties among them. Institution names were extracted from the collection of papers and disambiguated using the Derwent Analytics (v1.2) software product. Institutions were clustered into collaboration groups based on their co-occurrence in papers. A crossmap of clustered institutions against research fronts, which were derived using bibliographic coupling analysis, shows the research fronts that specific institutions participate in, their collaborator institutions and the research fronts in which those collaborations occurred. A crossmap of institutions to author teams, derived from co-authorship analysis, reveals research teams in the specialty and their general institutional affiliation, and further identifies the researchers that function as weak ties and the institutions that they link. This case study reveals that the techniques introduced in this paper can be used to extract a large amount of useful information about institutions participating in a research specialty.

Keywords

institutional collaboration; bibliometrics; mapping; visualization, cystic fibrosis body composition

Introduction

A research specialty is a self-organized social organization, which can be described by the different features associated with it: 1) a research paradigm, 2) knowledge structure, 3) personnel, 4) institutions, 5) specialized vocabulary, 6) collaboration structure, 7) research output, 8) and domain journals. Institutions provide the nurturing environment and infrastructure for researchers and play an important role in the evolution of a specialty. The mapping of institutional activity in a specialty will be helpful for decision making by policy makers and research funding agencies.

In previous examples of mapping institutions, research teams, derived from co-authorship patterns, are used to reflect the collaboration pattern among institutions (Nagpaul, 2002; Kretschmer *et al*, 2005; Havemann, *et al*, 2006). Katz and Martin (1997) described the distinction between collaboration and co-authorship, and pointed out that the two are not necessarily equivalent. Due to many problems experienced when extracting institutions, few researchers have attempted to map institutions from

¹ This work was partially supported by the Ministry of Science and Technology of China (Grant 2004CCC00400).

² The author thanks Professor Jin Bihui of the Library of Chinese Academy of Sciences for her valuable help and encouragement.

journal literature. As a notable exception, Borner, *et al* (2006) described techniques for extracting and normalizing institution names based on co-citation links, and analysed the diffusion of scholarly knowledge among major U.S. research institutions.

In recent years, social network theory was introduced in bibliometric study (Leydesdorff, 1998; Debackere & Clarysse, 1998; Kretschmer & Aguillo, 2004). Network theory focuses on organizations' and authors' knowledge interaction and communication, and views social relationships in terms of nodes and ties (Wasserman, *et al*, 1994). Nodes are the individual actors within the networks, and ties are the relationships between the actors. Granovetter (1973) discriminated between 'strong' ties that occur among close co-workers, and 'weak' ties that occur between individuals that interact infrequently. He noticed that by using weak ties, individuals could communicate beyond their well defined social boundaries.

The concept of weak ties in social networks was extended to the study of scientific specialties by Chubin (1976), who conjectured that an important class of weak tie researchers consists of marginal workers with no incentive to pursue research within existing paradigms. In addition, Bjerneborn and Ingwersen (2001) noticed the weak tie phenomenon in webometrics perspectives. Wagner (2005) argued that weak ties, evidenced by geographically remote collaboration, can promote new knowledge creation.

Detecting and visualizing researchers that function as weak ties between institutions will be helpful for demonstrating communication of knowledge, and may further identify the structure of invisible colleges. Using the coauthorship network of a specialty, it is possible to use graph layout software such as Pajek (Wouter, *et al*, 2005) or UCINET³ to visualize author teams and weak ties between teams. On such a map, teams appear as closely co-located clumps of authors with dense intra-team link networks, while weak ties appear as authors at the ends of inter-team links. This visualization technique is difficult to use in practice because the denseness of author labels interferes with easy interpretation of the plots.

In this article, we will describe techniques to extract and clean up institutions in a collection of papers covering a specialty. The extracted institutions were used to build a paper to institution matrix for analysis and mapping of institutional collaboration. We apply timeline and crossmapping techniques to visualize the collaboration structure of institutions, show institutional links to individual researchers and research teams, and most importantly, show new techniques for visualizing weak ties in a specialty. We demonstrate these techniques through a case study of cystic fibrosis body composition, a medical specialty. A subject matter expert (SME) was used in our study to guide data collection and to validate analysis and visualization results.

Experimental Technique

Collection of data

The source data for this study was gathered on August 7, 2006 from ISI/Thompson's Web of Science (WOS) product, using the following topic query: ***"cystic fibrosis" AND ("nutritional status" OR "body composition" OR "fat mass" OR "fat free mass" OR "lean body mass" OR "body cell mass" OR "bone mineral content" OR "total body water" OR "bone mineral density" OR "body density" OR "dual energy x-ray absorptiometry" OR "bioelectrical impedance" OR "air displacement plethysmography" OR "anthropometry")***. The query was applied against publication years from 1900 to 2006. This query was developed by our SME to retrieve papers on the specialty from the PUBMED archive for an earlier contract study.

³ Analytic Technologies, P.O. Box 920089, Needham, MA 02492, USA

Description of data

The downloaded source data consists of Web of Science (WOS) records in ISI tagged file format, containing records for 622 papers, with 22257 citations to 12411 references, and 3164 authorships by 2032 authors. Publication dates of the papers range from 1978 to August, 2006.

Extraction of institutions

The task of extracting institutions from the WOS source file is critical for successful analysis of participation of institutions in a research specialty. To achieve the desired results, researchers face two main obstacles to successful extraction of institutions: 1) initial extraction of institution names, and 2) disambiguation of institution names. Note that in WOS files the authors are not positively associated with their institutions (Katz & Martin, 1997), so every author and his or her corresponding institution cannot be directly matched.

To extract and disambiguate institution names, we employed *Derwent Analytics*⁴. The software can input collections of papers in WOS tagged file format and transform that data into a database configuration file. Fields, such as author field, affiliation field, and journal field are extracted from the raw file using pattern matching, rule-based extraction, and natural language processing techniques. Our institution extraction process consisted of the following steps:

- 1) **Import of raw source file and extraction of institution name.** In the raw WOS record file, the institution name is obtained from the author affiliation address, which is tagged “C1” in the WOS format. In each record each author may have more than one affiliation, and furthermore, if two or more authors in the paper have the same affiliation, that affiliation is only listed once in the paper’s affiliation field. Derwent extracts one institution per C1 line as an ‘unnormalized’ institution. In our study, 558 unnormalized institutions were extracted from the source file.
- 2) **Disambiguation of institution names.** In many cases, different institutions might have the same name, or alternately, the same institution might have several different names. For accuracy, it is important to separate or combine such ambiguous institution names. Derwent Analytics relies on dictionary and matching rules to perform the combine function, but it cannot separate the different institutions that have the same name. In our case study, manual editing was required to correct 5 names (such as “Children’s Hosp”, “Hosp Sick Children” and so on) that Derwent Analytics could not correctly separate. This manual step required about 20 man-hours of labor.
- 3) **Producing an institution-paper matrix.** Based on the normalized institution list, Derwent Analytics was used to produce a paper to institution matrix. Institutions associated with five or more papers were retained for analysis, resulting in a matrix of 50 institutions to 622 papers.

Co-occurrence and similarity computation

Data was loaded into three matrices for analysis: 1) a paper to reference matrix, 2) a paper to paper author matrix, and 3) a paper to institution matrix. Cooccurrence counts were computed and converted to similarities using the Dice formula (Salton, 1989). Distances for clustering were computed by subtracting similarities from unity. All clustering analysis was done using hierarchical agglomerative clustering with Ward’s method linkage. Three types of co-occurrence clustering were applied to the data:

- 1) **Clustering of papers by common references**, that is, bibliographic coupling clustering (Morris, *et al*, 2003). This type of analysis produces groups of papers that tend to use the same base knowledge, as represented by the common references that they cite. We shall refer to these groups of papers as ‘research fronts’ (Morris *et al*, 2003), although a longer and more appropriate term would be ‘bibliometric research fronts’. For this analysis, those papers not having 5 or more bibliographic coupling counts to at least one other paper were discarded, leaving 461 papers for clustering. The resulting clustering hierarchy was truncated to produce 20 research fronts.

⁴ Thomson Scientific, 3501 Market St., Philadelphia, PA., 19104, USA

- 2) **Clustering of paper authors by common paper**, that is, co-authorship clustering. This type of analysis tends to produce groups of authors corresponding to research teams (Melin & Persson, 1996). In this analysis, authors having less than 5 papers were discarded, leaving 83 highly productive authors for analysis.
- 3) **Clustering of institutions by common paper**. This type of analysis tends to produce groups of institutions that collaborate on common research projects, or that have researchers with affiliations in both institutions, as will be explained below. In this analysis, institutions having less than 5 papers were discarded, leaving 50 highly productive institutions for analysis.

Visualization

Visualization of research front structure, institution collaboration structure and author collaboration structure was conducted using timeline techniques (Morris, *et al*, 2003) and crossmapping techniques (Morris & Yen, 2004). These are matrix-based techniques that display the matrix of occurrence counts between entities of two different entity-types in the collection.

Figure 1. shows a timeline of research fronts in the collection of papers. Papers are plotted on the y-axis in horizontal tracks by research front, and on the x axis by publication date. The research front clustering dendrogram and research front numbers are shown on the left side of the plot, while the research front labels appear on the right side. Papers are plotted as circles whose size is proportional to the number of times the paper has been cited, and whose shading is proportional to the number of times the paper was cited in the final year of the collection. This allows easy identification of important papers, and currently ‘hot’ topics in the collection.

Figure 2 shows a crossmap of research fronts to institutions in the collection. The research fronts, on the y-axis, have the positions, research front numbers, labels, and clustering dendrogram taken from the timeline of Figure 1. The clustering dendrogram and institution numbers, which are arbitrarily assigned during clustering, are shown at the top of the plot. Corresponding institution labels are shown along the bottom of the plot. Given institution i on the x-axis, and research front j on the y-axis, the size of the circle at position (i,j) is proportional to the number of papers in research front j that are associated with institution i through author affiliation. This plot was constructed to visualize the relation of institutions to research fronts, and by inference the sub-topics, in the specialty. The plot also gives indications of each institution’s research in terms of the number of research fronts it participates in, the productivity of each institution, and the collaboration structure of the institutions.

Figure 3 shows a crossmap of institutions against paper authors. The institution numbers, labels, and clustering dendrogram are the same as those from the y-axis of the research front to institution crossmap of Figure 2. The paper author numbers, arbitrarily assigned during clustering, and the clustering dendrogram appear above the plot. The corresponding author names appear below the plot. Given author i on the x-axis, and institution j on the y-axis, the size of the circle at (i, j) is proportional to the number of papers to which author i and institution j were both associated. This plot was constructed to give visual indications of the productivity of individual authors, their relation to specific institutions and the breadth of author participation in terms of the number of institution that author is associated with. The plot also indicates the basic collaboration structure among the authors, which shows clearly research teams in the specialty. We also use this plot to visualize weak ties among institutions in the specialty.

Case Study: Cystic Fibrosis Body Composition

Validation with SME

Our subject matter expert, the third author of this paper, is a medical researcher with recent experience in the specialty, who has recently completed a review of the specialty’s literature under contract from a university hospital. The SME is familiar with the specialty’s research topics, key papers, leading researchers and leading institutions.

The visualization and analysis results of this study were presented to the SME for assessment of conclusions against expert knowledge. The SME was provided with the three visualizations of this paper (Figures 1, 2 and 3) and an automated report which contained summary data for each research front:

1. A list of papers in the research front, including authors, title, journal, published year, volume, and page.
2. A list of references cited, ranked by number of citations received.
3. A list of paper authors, ranked by number of papers authored.

The SME was asked to provide background on the specialty and to assess the analysis and visualization results for accuracy.

Background

Cystic fibrosis (CF) is a commonly occurring inherited disorder characterized by chronic pulmonary inflammation and gastroenterological abnormalities. Many children with CF experience growth retardation, malnutrition, and premature mortality due to respiratory insufficiency. Severe growth delay and poor nutritional status in terms of reduced height for age, low body weight for age, and pronounced deficits in lean body mass are predictors of increased morbidity and mortality. Nutrition intervention studies have suggested that improvements in nutritional status may mitigate declines in pulmonary status, yet understanding of the etiology of malnutrition and its role in disease progression in CF remains incomplete.

Classic studies in the 1980s demonstrated that diet therapies which resulted in an increase in body weight were associated with improvements in growth, clinical status, and survival. These findings, coupled with subsequent research into the nature of energy requirements and negative energy balance as well as improvements in pharmacological options, have contributed to considerable advances in the nutritional and clinical management of patients with CF (Pencharz & Durie 2000), so that life expectancy has increased into the third decade of life.

Research topics regarding body composition in CF tend to be divided into subcategories:

- Many researchers restrict their investigations according to age (pediatric versus adult).
- Some investigators focus on a given technique and its utility in characterizing body composition in healthy individuals and in those with various diseases.
- Some investigators focus on a given population (such as subjects with CF) and examine body composition in that context using multiple techniques.
- Research may be further restricted to the study of certain body composition characteristics, such as bone mineralization and density, or abnormalities in the distribution of extracellular fluid, and their relationship with morbidity or outcome.

Description of research fronts

Because of limited space in this paper, we will only cover notable highlights in the results. Research fronts are displayed in the timeline of Figure 1. Using the abbreviation RF for 'research front', note that RF 17 and RF 18 are papers concerned with bone mineral density. This group shows a stairstep pattern of papers with many highly cited early papers (large circles), and currently highly cited papers (red circles). When using timeline techniques, this stairstep characteristic is typical for 'hot topic' research fronts that develop after some discovery (Morris & Boyack, 2005). In this case, a new method of measuring bone density, dual x-ray absorptiometry (DXA), was developed and became available in the early 1990's. Papers in RF 17 exploited this technique as well as earlier techniques that characterized bone health. Papers in RF 18 heavily cite papers in RF 17 and also heavily cite a 1999 study by Haworth which established by DXA that low bone density was common in a large group of patients with varying degrees of clinical severity in their lung disease. This finding, in conjunction with the accessibility of DXA and a growing population of adult patients with CF, has resulted in prolific research into bone health in patients with CF that is ongoing today.

Analysis of institutional participation in research fronts

Examining the research front to institution crossmap of Figure 2, the following interpretation is possible:

- Clustering and dendrogram seriation used in the crossmapping technique have placed the institutions into three general groups. The left half of the map, left to right from Institution 6 to Institution 13 corresponds generally to U. S. and Canadian institutions. The right half of the map, left to right from Institution 35 to Institution 21 corresponds generally to Australian and European institutions. The center, left to right from Institution 26 to Institution 46 corresponds to institutions that do not show evidence of collaborating with other institutions, that is, the institutions that have little similarity to other institutions based on co-authorship in papers. In our experience, such central placement of items with low similarity is a characteristic of the crossmapping technique.
- There is a distinct pattern of affiliation of universities with local hospitals, for example, the University of Pennsylvania, on the extreme left, is matched with co-located Children's Hospital of Philadelphia, which is located adjacent to U. of Pennsylvania on the crossmap. Several of these university-hospital close pairings are evident, particularly on the Australian/European side of the map. Our SME conjectures that these pairings reflect single researchers that are affiliated with both a hospital and a university medical school. An examination of the affiliation lists in selected papers generally verifies this conjecture.
- The following research fronts have dominant institutions:
 - RF 1, 'diagnosis and management' – U. of Wisconsin and associated hospital.
 - RF 15, 'eating behavior' – Brown U. and U. of Cincinnati
 - RF 2, 'nutritional status' – U. of Melbourne and associated hospital.
 - RF 5, 'growth hormone' – U. of Texas
 - RF 6, 'tissue depletion', U. of Wales
 - RF 9, 'calcium intake', U. of Cincinnati, U. of Florida, and Ohio State U.
 - RF 16, 'exercise', U. Utrecht
 - RF 20, 'energy expenditure', U. of Pennsylvania, U. of Queensland and associated hospital

Analysis of team participation in institutions and weak ties

The institution to author crossmap of Figure 3 is used to investigate participation of teams in institutions and to further indicate which authors function as weak ties among the institutions. In this figure the institutions appear on the y-axis using the same dendrogram and serial order as they appeared on the research front to institution crossmap of Figure 2. This allows authors that correspond to particular institutions in the author-institution crossmap to be related back to research topics through the research front-institution crossmap. The following indications can be derived from the institution-author crossmap:

- In the upper left, a group of Australian authors appear as two collaborating teams: left to right authors 81 to 47 form a team drawn from U. of Queensland, Queensland U. of Technology, and Brisbane Royal Childrens Hospital. The authors from Brisbane Royal Childrens Hospital also participate with other Australian hospitals such as Prince Charles Hospital, but do not collaborate with either of the Australian Universities when doing so.
- A distributed group of researchers appears at the bottom left center of the map, consisting of authors drawn from the Eastern U.S. These include researchers at U. of Pennsylvania, the Children's Hospital of Philadelphia, Brown University (Rhode Island), U. of Michigan, U. of Cincinnati, U. of Florida, Children's Hospital – Columbus (Ohio), and Ohio State U. (Columbus). Our SME conjectures that this may indicate a group of authors that have participated in a multi-center trial, where patients from several institutions are pooled to produce a study with a sufficiently large number of subjects. Results from such a study could be authored by researchers from all participating institutions.

Note in the institution-author crossmap that authors, which correspond to columns on the map, typically have a large circle that presumably corresponds to their home institution, and several smaller

circles that represent the institutions that the author has collaborated with. For example, Author 3, M. Corey, has a large circle corresponding to Toronto Hospital for Sick Children and also U. of Toronto, presumably the author's home institution. M. Corey also has collaboration links to McGill U., U. of Wisconsin, Ohio State U. (in Columbus), and Columbus Children's Hospital. This is a good indication that Corey functions as a weak tie between her institutions and the other institutions listed. Thus, from this characteristic of the institution-author crossmap, it can be used to visualize and identify weak tie authors and the institutions that they connect.

Note in the institution-author crossmap that authors 42, 4, 51, 12, and 17 form a series of links to several institutions across the lower right of the map: U. of Texas, U. of North Carolina, Cystic Fibrosis Foundation, Stanford U., and U. of Cincinnati. This system of links corresponds to coauthorship of these authors and institutions of "Consensus statement: Guide to bone health and disease in cystic fibrosis" published in *Journal of Clinical Endocrinology And Metabolism* in 2005. This indicates that these researchers function as weak ties among their institutions. Our SME conjectures that a system of professional ties probably existed among these researchers prior to writing the consensus paper.

Conclusion

We have shown that Derwent Analytics software can be used to extract institutions from ISI files. Using extracted institutions, timeline and crossmapping techniques were applied to visualize the institutions' participation in the specialty. Using the author to institution crossmap, we investigated the detection of weak ties and their institutional connections.

In our case study, the SME validated that we have found the main institutions in cystic fibrosis body composition specialty; and that the timeline technique correctly illustrated that bone mineral density research exploiting DXA is a 'hot topic.' The detailed information about the collaboration structure among these institutions and the research fronts they participated in was described. In addition, the weak tie authors and their institutions were observed directly. The resulting visualizations were examined and interpreted by our SME, demonstrating their readability and practical value for researchers, policy makers and funding agencies.

Extracting institutions is a critical step in our study, though Derwent Analysis can do this work easily, it can not normalize institution names as well as we anticipated, particularly, it cannot distinguish the different institutions that have the same name and different addresses. In our study, we had to expend 20 hours of manual work to separate them, a process prone to manual errors.

The study presented here illustrates the need for 'smarter' institution extraction software. While the use of the crossmapping techniques for finding weak ties appears to be theoretically well-founded, the great abundance of researchers with multiple affiliations appears to have caused interpretational ambiguities, a fact that was immediately apparent to our SME when evaluating the visualizations. What is necessary is the development of institution extraction software that can reliably connect authors to their affiliations. This software will be difficult to develop, but will allow significant advances in bibliometric analysis of institutional collaboration in specialties.

References

- Bjorneborn, L., & Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50(1), 65-82.
- Borner, K., Penumathy, S., Meiss, M., & Ke, W. M. (2006). Mapping the diffusion of scholarly knowledge among major us research institutions. *Scientometrics*, 68(3), 415.
- Chubin, D. E. (1976). Conceptualization of scientific specialties. *Sociological Quarterly*, 17(4), 448-476.
- Debackere, K., & Clarysse, B. (1998). Advanced bibliometric methods to model the relationship between entry behavior and networking in emerging technological communities. *Journal of The American Society for Information Science*, 49(1), 49.
- Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 778(6), 1360-1380.
- Havemann, F., Heinz, M., & Kretschmer, H. (2006). Collaboration and distances between German immunological institutes. *Journal of Biomedical Discovery and Collaboration*, 1(1), 6-6.
- Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, 26(1997), 1-18.

- Kretschmer, H., & Aguillo, I. F. (2004). Visibility of collaboration on the web. *Scientometrics*, 61(3), 405.
- Kretschmer, H., Kretschmer, U., & Kretschmer, T. (2005). Visibility of collaboration between immunology institutions on the web including aspects of gender studies. *Paper presented at the 10th International Conference of the International Society for Scientometrics and Informetrics* (pp.750-760). Stockholm, Sweden : Karolinska University Press.
- Leydesdorff, L. (1998). Theories of citation? *Scientometrics*, 43(1), 5-25.
- Melin, G., & Persson, O. (1996). Studying research collaboration using co-authorships. *Scientometrics*, 36(3), 363-377.
- Morris, S. A., & Boyack, K. W. (2005). Visualizing 60 years of anthrax research, *10th International Conference of the International Society for Scientometrics and Informetrics* (pp. 45-55). Stockholm, Sweden: Karolinska University Press.
- Morris, S. A., & Yen, G. (2004). Crossmaps: Visualization of overlapping relationships in collections of journal papers. *Proceedings of the National Academy of Sciences of the United States*, 101(suppl. 1), 5291-5296.
- Morris, S. A., Yen, G., Wu, Z., & Asnake, B. (2003). Time line visualization of research fronts. *Journal of the American Society for Information Science and Technology*, 54(5), 413-422.
- Nagpaul, P. S. (2002). Visualizing cooperation networks of elite institutions in India. *Scientometrics*, 54(2), 213.
- Pencharz PB, Durie PR (2000). Pathogenesis of malnutrition in cystic fibrosis, and its treatment. *Clinical Nutrition*, 19(6), 387-394.
- Salton, G. (1989), *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Compute.*, Reading, MA:Addison-Wesley, 1989
- Wagner, C. S. (2005). Six case studies of international collaboration in science. *Scientometrics*, 62(1), 3-26.
- Wasserman, S., Faust, K., & Iacobucci, D. (1994). *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press.
- Wouter, D.N., Andrej, M. & Batagelj V. (2005) *Exploratory Social Network Analysis with Pajek*. New York,Cambridge University Press



Figure 1. Timeline of research fronts in the CF body composition data set.

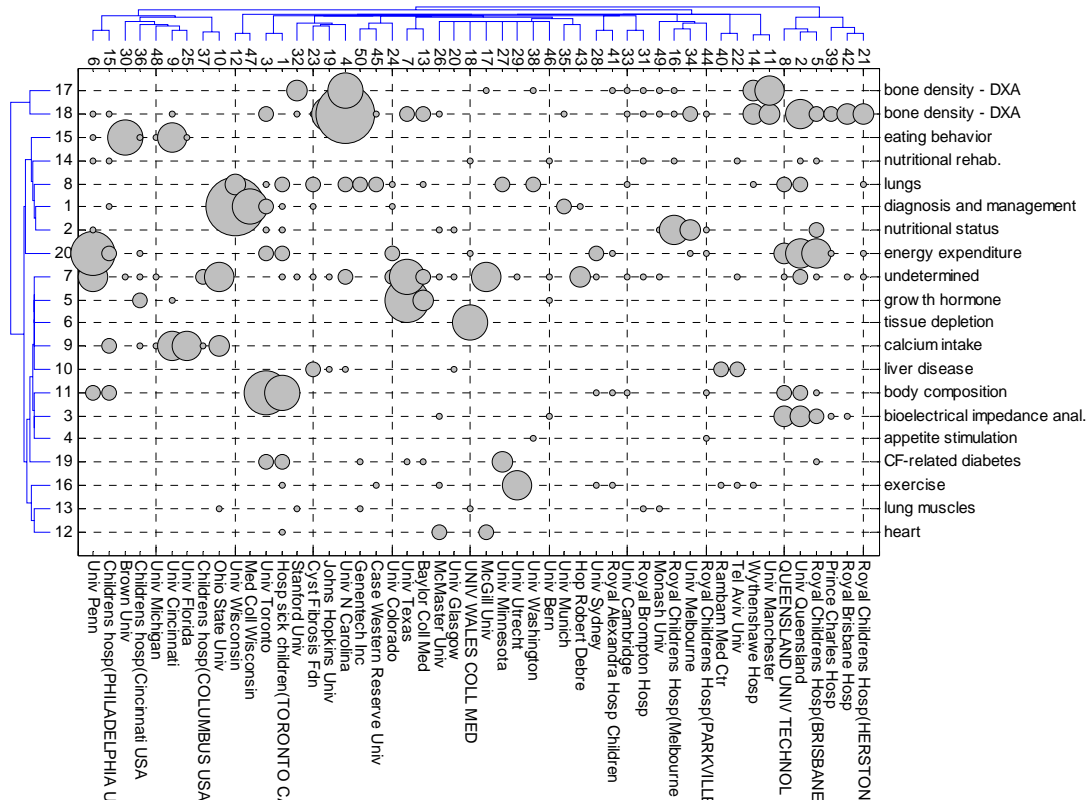


Figure 2. Crossmap of research fronts against institutions in the CF body composition data set.

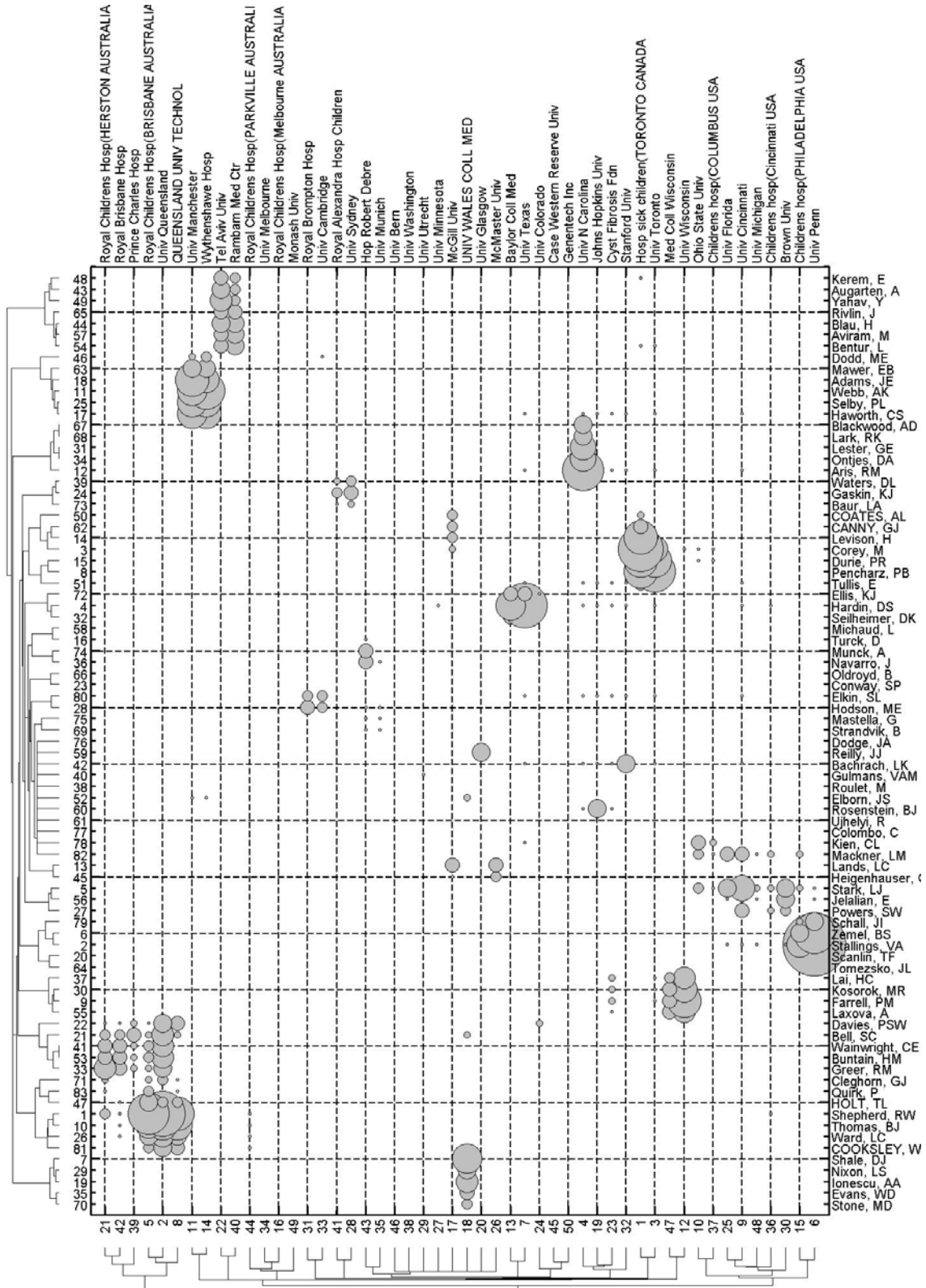


Figure 3. Crossmap of institutions against authors in the CF body composition dataset.