

Mapping Scientific Disciplines and Author Expertise Based on Personal Bibliography Files

Colin Murray+, Weimao Ke*, Katy Börner*

+University of Sydney, NSW 2006, Australia

+National ICT Australia, Locked Bag 9013 Alexandria, NSW 1435 Australia

*School of Library and Information Science, Indiana University

10th Street & Jordan Avenue, Bloomington, IN 47405, USA

cmurray@it.usyd.edu.au, wke@indiana.edu, katy@indiana.edu }

Abstract

This paper presents a novel approach (1) to define, analyze, and map a scientific discipline and (2) to compare and map the expertise of single authors based on personal bibliography files, e.g., bibtex or EndNote files. Section one motivates this research and relates it to existing work. Section two explains the general procedure from data harvesting, parsing, cleaning via the analysis and mapping of the data to the interpretation of results. To illustrate the new approach, we asked major experts on 'network science' to share their bibtex or EndNote files with us. Using this data set we exemplify data harvesting, data cleaning and integration; give simple statistics, analysis and visualization results such as a map of the network science discipline based on co-author networks and a map of the content coverage and overlay of contributing experts. We conclude with a discussion of challenges and opportunities as well as planned future work.

Keywords--bibliography, co-author network, network analysis, information visualization.

1. Introduction

Bibliometric and scientometric studies of scientific disciplines and the rendering of results as knowledge domain visualizations [3, 4, 17] aim to answer questions such as: What are the major research areas, experts, institutions, regions, nations, grants, publications, journals in a specific research field? Which areas are most insular? What are the main connections for each area? What is the relative speed of areas? Which areas are the most dynamic/static? What new research areas are evolving? Impact of the research in this field on other fields? How does funding influence the number and quality of publications? The studies are commonly conducted based on publication datasets downloaded from major digital libraries or online sources. In order to map a specific discipline, keyword based searches for relevant phrases are run or cited reference search is used to retrieve all papers that are cited by or are citing a set of seminal papers.

There are two major problems with this approach. Firstly, very few individuals have access to high quality publication data, e.g., the *Web of Science* served by Thomson Scientific. Secondly, at a time of increasing disciplinary specialization, it is very hard if not impossible to identify appropriate search phrases or the complete set of seminal papers that can be used to retrieve all relevant papers. This is particularly an issue for newly emerging or highly interdisciplinary research areas.

In this paper, we propose a novel approach to acquire a dataset that covers a specific area of research. Instead of querying databases, major experts in an area of interest are invited to share their personal bibliography files and to identify other experts that should be invited to submit personal bibliographies. Based on the resulting bibliography data, (1) an analysis and visualization of a scientific discipline can be conducted and (2) the content coverage and overlap of personal bibliographies can be compared and mapped.

Subsequently, we detail the data acquisition and data cleaning phase, describe the data analysis and mapping, and discuss results. The emerging, interdisciplinary area of 'network science' was chosen to illustrate the different steps and to interpret results. Obviously, the same approach can be applied to delineate and map other areas of science.

2. Data Collection

As mentioned before, our approach does not require access to any database. Instead we invited major experts to share their personal bibliographies with us. Interested to map the area of 'network science' we queried and received bibliography files from major experts in this area such as:

- Albert-László Barabási – a physicist most well-known for his research on scale-free networks. He is the author of 'Linked: How Everything Is Connected to Everything Else and What It Means' [2].
- Noshir S. Contractor – a researcher in the area of communication science. He co-authored 'Theories of Communication Networks' [9].
- Loet Leydesdorff – has published extensively in the philosophy of science, social network analysis,

scientometrics, and the sociology of innovation. He co-authored ‘The Challenge of Scientometrics: The Development, Measurement, and Self-Organization of Scientific Communications’ and ‘A Sociological Theory of Communication: The Self-Organization of the Knowledge-Based Society’ [7, 8].

- José F. F. Mendes – a physicist who co-authored ‘Evolution of Networks: From Biological Nets to the Internet and WWW’ [6].
- Mark E. J. Newman – a major expert in the area of physics & social sciences [10-12].
- Mike Thelwall – a researcher in the area of information science, particularly webometrics. He authored ‘Link Analysis: An Information Science Approach’ [14].
- Alessandro Vespignani – a physicist studying the Internet as well as the spread of epidemics. He co-authored ‘Evolution and Structure of the Internet: A Statistical Physics Approach’ [13].
- Duncan Watts – a researcher in the area of sociology and author of ‘Small Worlds: The Dynamics of Networks between Order and Randomness’ [16].
- Stanley Wasserman – a major social network researcher and statistician and co-author of ‘Social Network Analysis: Methods and Applications’ [15].

Our request for data used the following wording:

”Dear xxx,
 László [Barabási] inspired an exercise to map Network Science on a large scale. Stan [Wasserman], László, Alex [Vespignani], and I decided that the bibliographies of major books & review articles might be a good starting point.
 I would appreciate if you point me to (don't send me the MB files!) as clean and as complete as possible bibliographies. EndNote or bibtex file format will work best. We will then parse out and map the co-author network.
 Ideally, the resulting map will become an effective visual interface to major papers, books, etc. and expertise. It might also help to understand the structure and evolution of this growing field. The map will become particularly valuable if network science researchers start to enter their own publication data as it becomes available.
 Best regards,
 k

PS: Should I contact anybody else to contribute bibliographies?”

In response to our request, we received 13 files in EndNote, bibtex, and free text format that contained over 7,000 articles including duplicates. Subsequently, we explain the data cleaning, analysis, and visualization.

3. Data Parsing and Data Integration

Bibliography files in standard format such as EndNote & bibtex required the coding of special parsers but did not pose a problem in terms of automatic processing. Parsers were written to extract data from

EndNote formatted files. CiteULike (<http://www.citeulike.org/>) was used to convert bibtex files into EndNote format. Some bibliography files were provided in free form text – analogous to the lists of references one sees in publications or books. Diverse approaches were tried to parse and clean this data yet the result quality was unsatisfactory due to inconsistencies in the formatting. Some authors provided files in different formats. Ultimately, Wasserman’s and Leydesdorff’s files, about 50% of Mendes’ and Newman’s files, and a very small percentage of Barabási’s files could not be parsed automatically and had to be excluded from the subsequent analysis.

The resulting cleaned bibliography database comprised 5,425 unique articles (match based on title information) and 5,330 unique authors (based on last name and first initial) in the area of network science across several disciplines including physics, biology, information science, and social sciences. Because each set of articles had been contributed by a certain expert, each unique article can be associated with all expert contributors that listed/submitted this article. Hence, the coverage and overlap of personal bibliographies can be studied. Figure 1 shows the database schema used to store the bibliographic data.

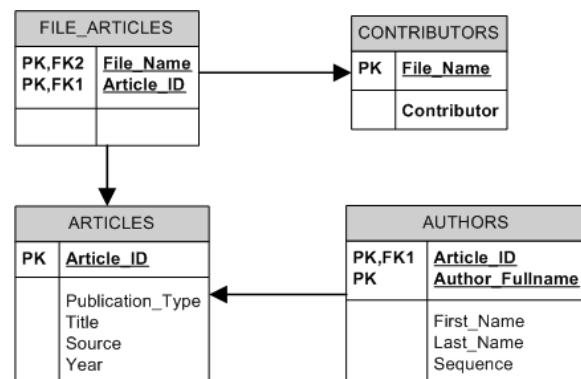


Figure 1: Database schema interlinking contributor, article, and author information.

It quickly became obvious that most contributors had generously sent us their complete bibliography files – comprising not only the references in their books but potentially all papers they ever read and cited. It also appears to be common – and is quite convenient – to share and merge bibliographies with co-authors. Co-authors with a long joint history or co-authors that merged large bibliographies might be easily identifiable.

Unfortunately, researchers use EndNote fields quite creatively. ‘Publisher’ entries end up in ‘pages’ fields, there are some cases of an article having no ‘title’ entry. While asking major researchers for the bibliography files is an easy way to quickly acquire data, getting that data cleaned and into a database is difficult due to formatting and consistency problems. Parsing and data integration problems and the need to exclude some files from the analysis are likely to adversely affect the subsequently presented results.

4. General Statistics

With the cleaned data in a properly setup database, diverse data analysis queries can be run easily. The oldest article was published in 1637 and was contributed by Contractor. The year with the maximum number of articles is 2002. Figure 2 shows the distribution of articles per author in a log-log plot. It clearly follows a power law with very few authors having published many papers while most authors have very few papers.

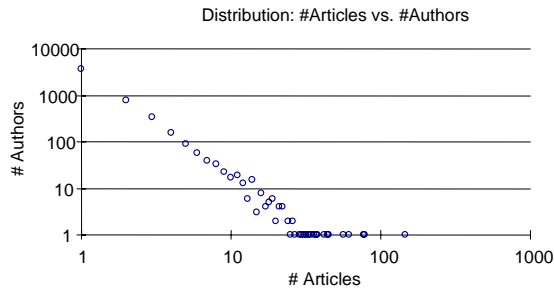


Figure 2: Number of articles per authors in log-log plot.

Next, we were interested to see how many bibliographic articles were supplied by the different contributors and when those articles were published. Figure 3 shows the distributions for the publication years 1975-2005. Watts and Barabási appear to have many recent articles in their bibliography files. Contractor also supplied a large number of articles yet they are more spread out across the years.

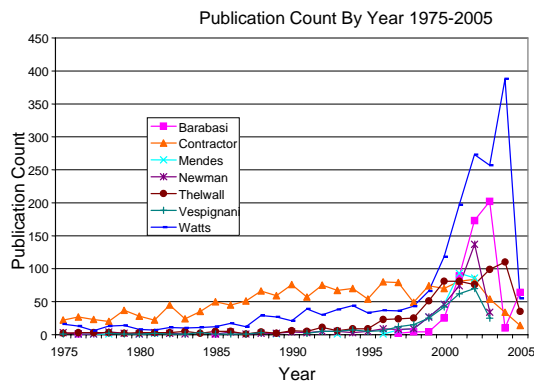


Figure 3: Number of contributed articles per publication year for all seven contributing experts.

5. Data Analysis and Visualization

Diverse exporting routines were written to export data in a format suitable for different data analysis and visualization tools such as the InfoVis Cyberinfrastructure (<http://iv.slis.indiana.edu/>), Gemi [1] a graph analysis and visualization tool that can be used to layout graphs in 2D or 3D, and Pajek [5] an analysis and drawing tools for large networks.

5.1 Mapping Scientific Disciplines Based on Personal Bibliographies

To map major experts in network science, we analyzed the co-author network of 5,330 unique authors using social network analysis [15] techniques. The authors are grouped into 266 components. The distribution of the component size is given in Figure 4. There are 14 components that have more or equal than 10 nodes. The largest component has 131 nodes.

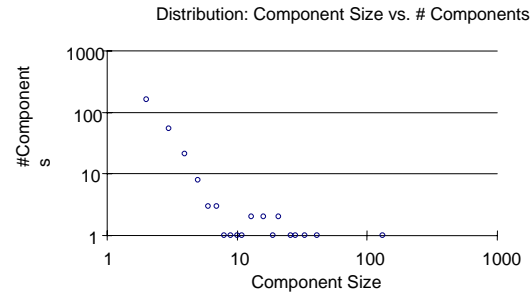


Figure 4: Component size vs. number of components in log-log plot.

The node with the highest degree denotes Jeong, who co-authored with 25 other authors in this data set. The co-authorship with the highest weight denoting a large number of co-authorships is between Dorogovtsev and Mendes who have co-authored 29 papers in this data set. Figure 5 shows the degree distribution of the network.

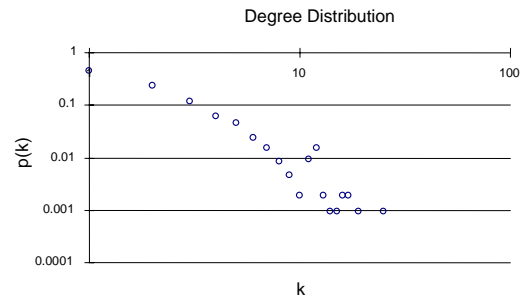


Figure 5: Degree distribution of the coauthor network in log-log plot.

Figure 6 (left) shows a Geomi force-directed, 2-D single layer layout of all authors with two or more articles and their co-authorships. The size of a node is based on the number of articles by that author. Its color denotes the node degree – red for nodes with degree 10 and higher, green for 5-9 links, blue for 2-4 links, and black for one or zero degree nodes. The width of the edge is based on the number of co-authorships. Nodes for authors with article count 10 or higher are labeled with the author's name. The author with the most papers is Leydesdorff. Assuming that all contributors submitted a large majority if not all of their own articles, this is particularly impressive as Leydesdorff's free text bibliography was not included due to parsing problems. Other authors with a large number of articles include two of the contributors: Barabási and Newman.

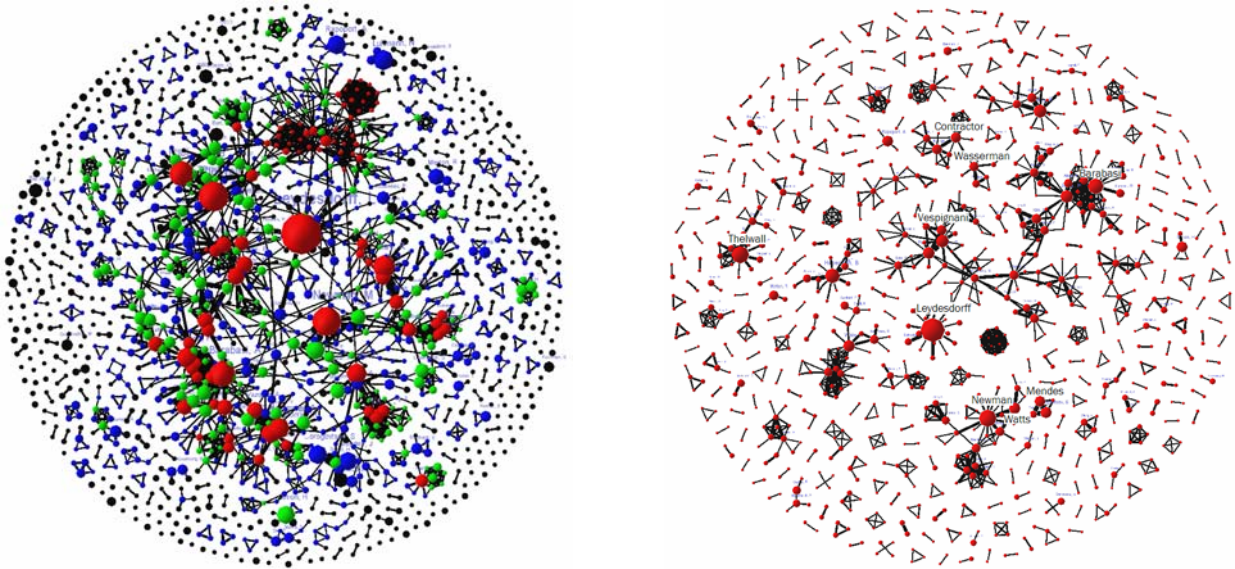


Figure 6: Co-author network of authors with two or more articles (left) and co-authorship network of authors who have written at least 2 papers together (right).

As expected, there is a correlation between the number of papers written and the number of co-authorships. However, there are also many authors that write papers with little or no collaboration. These are identified by the big black nodes. There are also cases of authors who have written few papers with several authors. These are identified by the high degree nodes that are small such as the red clusters towards the top of Figure 6 (left).

Figure 6 (right) shows the co-authorship network of all those authors that have written at least two papers together. As a result, 1,064 authors are shown. For a node to have a label in the visualization, the author must have published at least 10 papers. Exactly 129 scholars make this threshold. In this visualization, node size is used to indicate the number of articles and edge thickness indicates the number of papers coauthored.

Figure 7 shows the complete co-author network with triads and diads removed rendered in Pajek. Each component is given in a different color. All 579 authors are shown. Nodes with high BC value are denoted by a black outer ring. The giant component is colored in purple.

Figure 8 shows the 131 authors in the giant component of the co-authorship network using the same node positions as in Figure 7. Nodes and edges are colored based on time. See legend for details.

5.2 Comparing and Mapping Experts Based on Personal Bibliographies Expert Network

Given the personal bibliography files of seven major expert contributors, we were interested to analyze and map the coverage and semantic overlap of these experts based in the number of shared articles.

Figure 9 shows a network of all seven experts. Each node represents one of the contributors, labeled by its name and the number of articles supplied in parentheses. The size of each node corresponds to the number of articles contributed by this expert.

There is an edge between two experts if there is at least one article common to both experts. As for this dataset we have a fully connected network. The width of an edge denotes the number of articles common to both experts. The network was laid out in a 2D circular layout using Pajek.

As can be seen, Watts and Contractor contributed the most articles via their bibliography files. Watts and Barabási have the most articles in common while Contractor and Thelwall have the least in common with all other authors.

We hope this new approach to mapping scientific domains but also to mapping the coverage and overlap of experts will be widely adopted such that expert network visualizations can be generated for more than seven nodes. It might be truly exciting to see the network of all of science mapped this way.

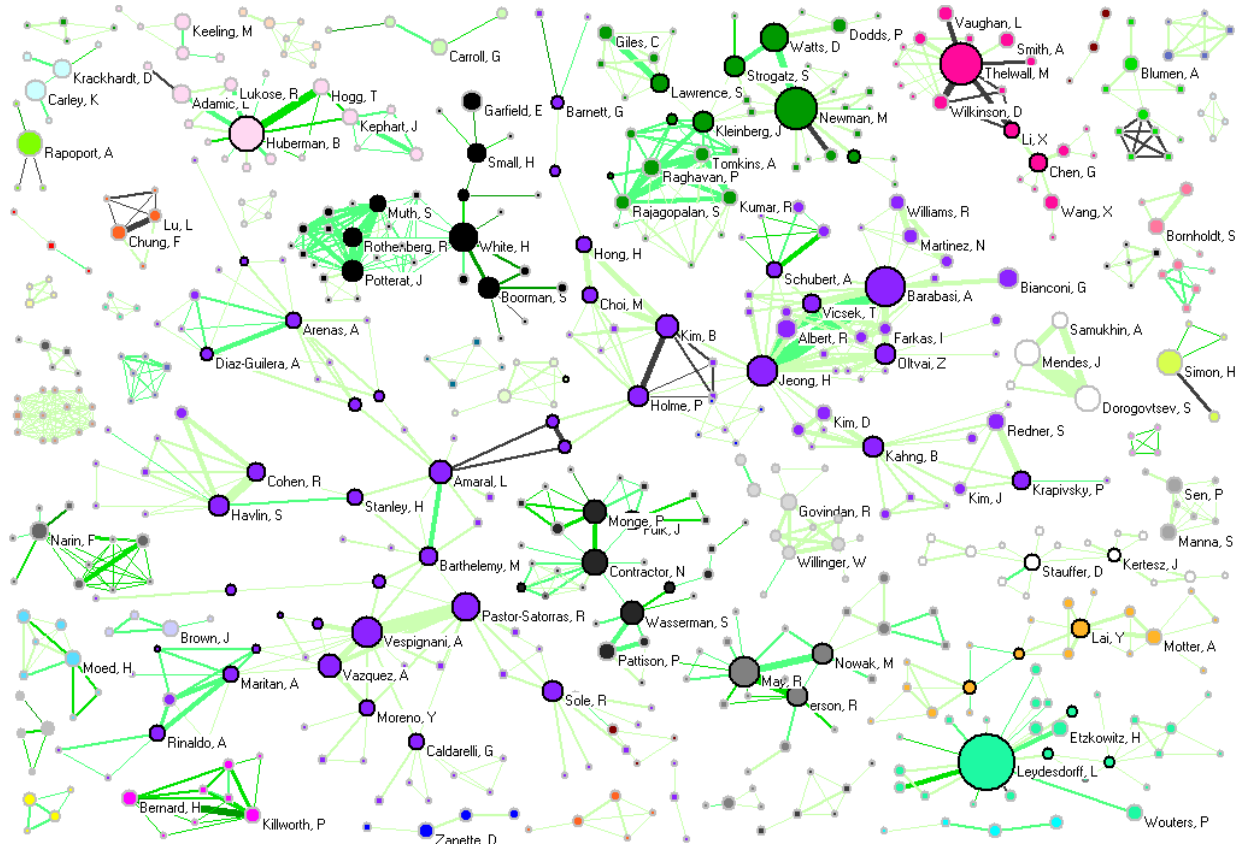


Figure 7: All components of the complete co-author network with size larger than three.

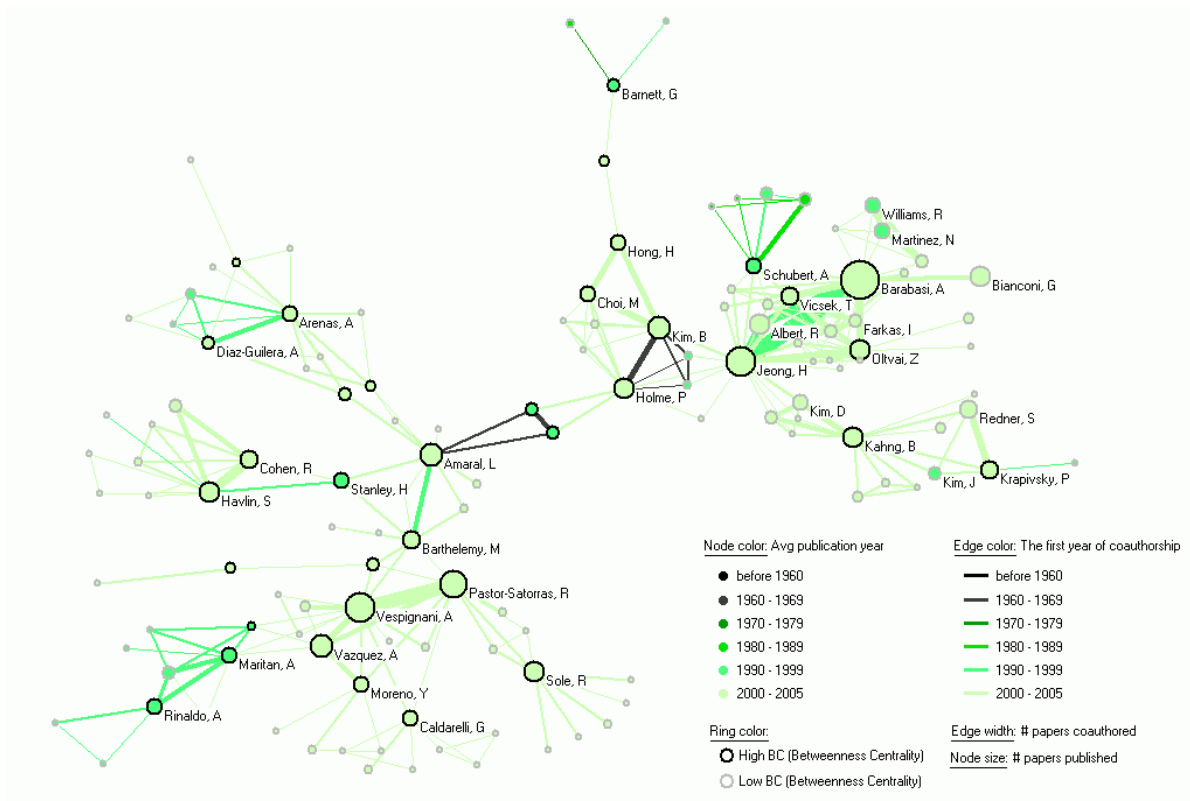


Figure 8: Giant component of the complete co-author network.

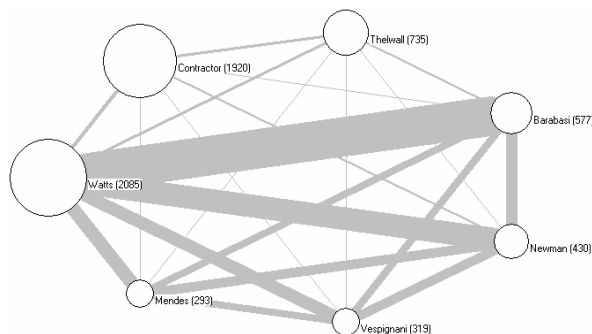


Figure 9: Expert network

Conclusions

The paper introduced a novel way of identifying a comprehensive collection of articles from a field of science based on personal bibliography data. The technique was demonstrated in the highly interdisciplinary field of 'network science' resulting in maps of network science authors and experts.

The presented technique can easily be applied to other domains of science. It has proven to be a fast way of collecting large amounts of bibliography data. In fact, it could provide a viable means to collect data required for analyzing and mapping all of science.

Obviously, the collected data does not provide any information on paper citation linkages. However, perfect bibliography data – supplied by hundreds and thousands of experts and integrated using automatic techniques – constitutes great raw material to query other databases such as Scholar.Google, Citeseer, etc. for citation linkages and citation counts.

We are in the process of setting up a web portal interface at <http://scimaps.org> that experts from diverse communities can use to upload personal bibliography files. In return, they will be able to download all unique articles submitted so far in EndNote, ISI, or bibtex format; retrieve simple statistics such as the number of articles per year or the top ten authors, papers, or journals; as well as co-authorship maps based on shared bibliographic entries.

Acknowledgements

We would like to thank László Barabási, Stanley Wasserman and Alessandro Vespignani for inspiring this work. Thanks also go to Albert-László, Noshir S. Contractor, José F. F. Mendes, Mark E. J. Newman, Loet Leydesdorff, Mike Thelwall, Alessandro Vespignani, Duncan Watts, and Stanley Wasserman for sharing their personal bibliography files with us.

This his research is supported by the National Science Foundation under IIS-0513650, CHE-0524661, and a CAREER Grant IIS-0238261 as well as a James S. McDonnell Foundation grant in the area Studying Complex Systems.

References

- [1] Ahmed, A., et al. (2005) GEOMI: GEOMETRY for Maximum Insight. in *13th International Symposium on Graph Drawing (GD 2005)*.
- [2] Barabasi, A.-L. (2003) *Linked: How Everything Is Connected to Everything Else and What It Means*.
- [3] Börner, K., Chen, C. and Boyack, K. (2003). Visualizing Knowledge Domains. in Cronin, B. ed. *Annual Review of Information Science & Technology*, Information Today, Inc./American Society for Information Science and Technology, Medford, NJ, 179-255.
- [4] Cronin, B. and Atkins, H.B.E. (2000) *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield*. ASIST.
- [5] de Nooy, W., Mrvar, A. and Batagelj, V. (2005) *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, Cambridge.
- [6] Dorogovstev, S.N. and Mendes, J.F.F. (2003) *Evolution of Networks*. Oxford University Press, Oxford.
- [7] Leydesdorff, L. (2001) *The Challenge of Scientometrics: The Development, Measurement, and Self-Organization of Scientific Communications*. Universal Publishers.
- [8] Leydesdorff, L. (2001) *A Sociological Theory of Communication: The Self-Organization of the Knowledge-Based Society*. Universal Publishers.
- [9] Monge, P.R. and Contractor, N. (2003) *Theories of Communication Networks*. Oxford University Press, New York.
- [10] Newman, M.E.J. (2001) Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, 64. 016131.
- [11] Newman, M.E.J. (2001) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64. 016132.
- [12] Newman, M.E.J. (2001) The structure of scientific network collaborations. in *Proceedings of the National Academy of Sciences*, 404 - 409.
- [13] Pastor-Satorras, R. and Vespignani, A. (2004) *Evolution and Structure of the Internet: A statistical physics approach*. Cambridge University Press, Cambridge.
- [14] Thelwall, M. (2004) *Link Analysis: An Information Science Approach*. Academic Press.
- [15] Wasserman, S. and Faust, K. (1994) *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences, 8)*. Cambridge University Press.
- [16] Watts, D.J. (1999) *Small World*. Princeton University Press, Princeton, NJ.
- [17] White, H.D. and McCain, K.W. (1989). Bibliometrics. in Williams, M.E. ed. *Annual review on information science and technology. Volume 24*, Elsevier Science Publishers, Amsterdam, Netherlands, 119-186.