Visualizing Networks of Innovation

Submission for the PCF 2007-2008
Awarded July 1, 2007

Visualizing Networks of Innovation

Investigators: Tim Lenoir (PI), John Madden, Cathy Davidson
Senior Personnel: Rachael Brady

1. Project Description

Knowledge of how to support the growth and diffusion of scientific knowledge and technological innovation is critical for creating effective environments for education, research, and economic development. The ecology of social, knowledge, and technological networks significantly affects who has access to them, both topically and temporally. A key problem of interest to federal policy makers, university research administrators, and regional planners is the role and impact of federal funding of scientific research in the stimulation of economic growth. Aggregate input-output scenarios combined with anecdotal evidence are most frequently used in such discussions. However, we live in an era in which rapid formation of flexible transdisciplinary collaborations involving close interaction among numerous scientific and engineering areas of research in close working relationships with industry have become critical to both scientific and industrial development. Hence, it becomes imperative to understand the local conditions and optimal configurations favoring successful interdisciplinary work in academic research settings and the transfer of knowledge to industry. Knowledge and technology diffusion in a dynamically evolving ecology of networks (e.g., social, co-author, paper-citation, patent, and funding networks) are complex, and do not yield their secrets to a single methodology.

The proposed project will address this issue by linking sophisticated analytical tools for identifying and clustering closely related documents with visualization techniques that enable users to understand complex diffusion processes. The visualizations, called “Knowledge Domain Visualizations (KDVs)” aim to communicate the results of the data analyses and to support the interpretation, discovery, understanding, and management of complex data sets. Knowledge domain visualizations are a special kind of *Information Visualization* that exploit powerful human vision and spatial cognition to help humans mentally organize and electronically access and manage large, complex information spaces.[4, 5] Unlike scientific visualizations, KDVs are created from data that have no spatial reference, such as papers, patents, and grants stored in digital libraries. KDVs use sophisticated data analysis and visualization techniques to objectively identify major research areas, experts, institutions, grants, papers, journals, etc., in a domain of interest. They can be used to gain an overview of a knowledge domain; to study its homogeneity, import-export factors, and relative speed; to track the emergence and evolution of topics; or to help identify the most productive as well as new research areas. Three sample visualizations are shown in Figure 1.

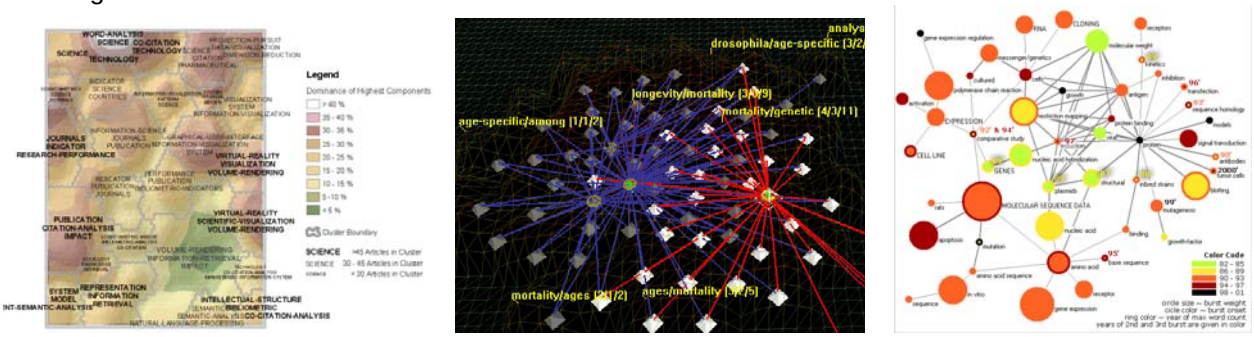

Figure 1. Visualization of research on visualizing knowledge domains using Self Organizing Maps and GIS by Skupin[14] (left); VxInsight map of correlated papers and grants in aging research by Börner[3] (middle); a map of the top 50 "hot" words in the most highly cited PNAS articles from 1982-2001 by Mane and Börner[10] (right).

Benefits of KDVs include reducing visual search time, revealing hidden relations, displaying data sets from several perspectives simultaneously, facilitating hypothesis formulation, serving as effective means of communication, and prompting users to think in new ways about document data.

To our knowledge, no one has yet attempted to study knowledge and technology diffusion in knowledge ecologies that have explicit time, geospatial and topic attributes. Moreover, visualization strategies, such as those depicted above, have yet to explore the third dimension. The visualizations we propose to develop will show the temporal, geospatial, and topical structure, as well as the evolution of diverse data elements. A key feature of our work will be the development of 3D visualization strategies. For example, we will show the structure and evolution of scientific domains, the migration of authors, and the diffusion of knowledge and technologies over geospatial

space and scientific domains. Geographic information system (GIS) software will be applied to handle complex, large-scale data sets and to arrive at visualizations that resemble cartographic maps and employ skills used when exploring geographic maps. GIS and geostatistical analysis will be used to perform neighborhood operations, to visually identify correlations by superimposing layers of information (e.g., funding information overlaid with output in terms of papers, new authors), and to visualize the diffusion of knowledge in both geospatial and topic space. These patterns can be quantitatively compared across different topic areas to compare how the spatial distribution of information flows differs by discipline or other criteria. Likewise, multi-temporal data can be viewed via spatial animations allowing the change in patterns and flows to be visualized. GIS functionality will be used to link to the database of network information. Geostatistical analysis will be conducted using point pattern analysis techniques.

For the purposes of this project we will concentrate in our first phase (Phase 1) of development on one particularly rich source of information, the U.S. Patent Database. This part of our project grows out of work that Tim Lenoir and Kevin Webb, with support from Hewlett-Packard and the Duke Computer Science Department, have been doing on extracting deep levels of information from patents. Patents offer significant sources of information for scholars interested in investigating the history of invention, the geographical location and diffusion of innovation. Unfortunately, the rich potential of information contained in patent documents is difficult to extract due to complex and historically shifting taxonomies employed in classifying inventions. Although citations of earlier patents by later patents help to generate lineages of directly related inventions, much of what is important in turning an invention into an innovation is more contextual and not captured by direct citation. Other data sets, frequently interdisciplinary and even transdisciplinary in character, including scientific and engineering literature outside items immediately cited by a patent document, and even other data sources, such as federal and private funding, need to be integrated into the account. For the proposed work, we assume that information diffusion occurs both directly via co-authorships and indirectly via production and consumption of the products of science such as papers. Hence, co-authorship relations and citation links become the major pathways of knowledge diffusion within a delicate ecology of networks of co-authors, paper-citations, and grants.

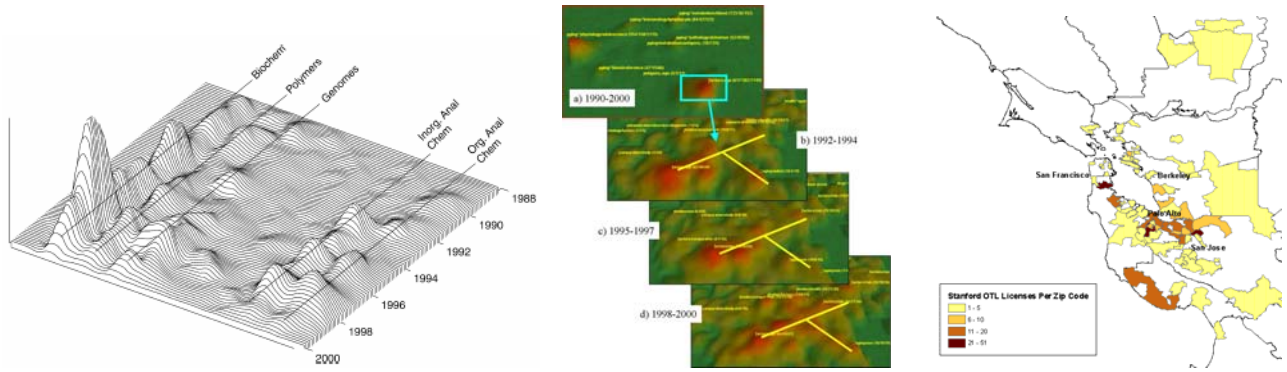

Figure 2. Topic by time plot of chemical sensor articles by Morris[10] (left); timeline of elderly-related aging papers by Boyack and Börner[3] visualized using VxInsight (middle), and Lenoir's[9] geospatial map of Stanford OTL licenses per zip code (right).

The complexity and multidimensional character of the knowledge diffusion networks described above can best be understood and interpreted with the aid of visualization tools. Visualizations of flow fields or the diffusion of substances over space and time are commonly used in physics or biology. Very few visualizations exist that show the diffusion of knowledge, and none of these access the third dimension. ThemeRiver maps the thematic strength of temporally collocated documents to the width of an imaginary 'river' through time.[8] Themes or topics are represented as colored 'currents' flowing within the river. Currents narrow or widen to indicate decreases or increases in the strength of a topic in associated documents at a specific point in time. Morris et al. developed a visualization system for exploring document databases for technology forecasting (see Figure 2, left). Boyack and Börner[3] used a time series of knowledge domain visualizations to show the evolution of elderly-related aging research over time (see Figure 2, middle).

2. Project Goals

The primary goal of Phase 1 of our project is to improve access to patent data through improved patent database tools and visualization, to study the effectiveness of the tools provided, and to do the required work to get the entire patent database accessible using an interface that is usable by researchers who are searching for specific criteria. In essence, this part of the project creates the technological framework that enables novel research to be done that would be impossible otherwise. For the proposed work, we assume that information diffusion occurs both directly via co-authorships and indirectly via production and consumption of the products of science such as papers. Hence, co-authorship relations and citation links become the major pathways of knowledge diffusion within a delicate ecology of networks of co-authors, paper-citations, and grants. However, these are notoriously difficult to visualize in two-dimensions, since these networks are so dense. Therefore, the use of visualization infrastructure to project this data using *extrusion* is needed. Extrusion is a novel technique that is currently under development by Rachael Brady that allows a researcher to interactively "reach into" a two-dimensional visualization of a network and "pull out" networks of patents into three-dimensions. In three-dimensions, diffusion should be much more easily discovered.

In Phase 2 of our proposed study we plan to couple the data extraction and analytical tools used in Phase 1 with visualization tools that allow us to integrate, display, and correlate several categories of information flow and diffusion. Specifically, Lenoir and Madden will focus on the development of regional flows related to biotech and medical practice within the Research Triangle area. The outcome we hope to achieve is the ability to locate federally funded pockets of innovation within academe (whether in single departments or in interdepartmental collaborations) and track the diffusion and impact of their work within industry and ultimately into research and medical practice. In addition to co-authorships, citation patterns of patents and scientific literatures, Lenoir and Madden will draw upon OneSource and Factiva for detailed data on commercial firms and industry sectors in North Carolina. We also seek to understand the feedback loops between sectors of local industry, university research and biomedical practice.

Specifically, the Common Fund project will focus on four objectives, the first two in Phase 1 and the second two in Phase2:

- *The creation of a freely accessible and searchable database of all U.S. patents for research scholars.*
- *Research, develop and demonstrate a 3D network visualization tool (NetVR).*
- *Use the visualization-patent framework to produce cutting-edge research that examines the biomedical industries of North Carolina and study the role of institutions (including institutions of higher education) in fostering or impeding innovation using the visualization of patents.*
- *Host a research symposium on the visualization of patents targeted to the Duke community in Spring 2008.*

Goal 1: The creation of a freely accessible and searchable database of all U.S. patents for research scholars.

Research in patent visualization faces one main obstacle: the availability of data. This is caused by the inability for researchers to retrieve multiple patents fitting specific search criteria, such as "all patents between 1975-76 assigned to Xerox PARC" and "all patents containing the word genomics". Without the ability to search the database and retrieve files matching very specific criteria, further work in patents is prohibitively expensive. At best, researchers would be forced to pay per patent at commercial services or use free services that force them to download a single patent at a time. Tim Lenoir has been working with Kevin Webb of Tackle Design on improving the access and retrieval of patents (see http://search.allpatents.org). Although this website is still in the early stages, it provides a substantial improvement over the interface used by the U.S. Patent Office itself, as it provides full text search of patents, returns patent data in aggregate and provides high quality TIFF files of the drawing associated with each patent. The site currently hosts over three million of patents filed in the U.S. between 1836 and 1976. This common fund proposal seeks funds for a database server which can host the entire U.S. patent collection as well as funds to complete the interface development with Kevin Webb which will provide full text search and retrieval capabilities. Cathy Davidson will further ensure the infrastructure built by this project is accessible to other interdisciplinary research at Duke.

Goal 2: Research, develop and demonstrate a 3D network visualization tool (NetVR).

The visualization of complex networks is difficult at best[1]. Two-dimensional representation over-simplifies the ways networks are intertwined in multiple ways and do not allow for interactivity that allows the researcher to unravel complex clusters. State-of-the-art network visualization tools rely on multiple 2D windows onto the underlying data which support interactive brushing and linking[13]. Currently no research infrastructure exists that allows generic networks to be taken from a two-dimensional space and visualized *interactively* in three-dimensional space for purposes of data exploration and analysis. This innovative idea is the heart of the proposed visualization tool.

With the help of faculty involved in the Franklin Humanities Seminar on Interface, Brady, Davidson and Lenoir have begun brainstorming a case-study that will lead to the creation of the preliminary network visualization research software infrastructure, tentatively entitled NetVR (Network Virtual Reality Toolkit). For this initial study, Kevin Webb has retrieved all patents that contain keywords *Xerox Parc, Interactive Computing, Personal Computer Interface, Internet, Virtual Reality, Licklider, and Engelbart;* over 10,000 patents. By March 2007, the Brady lab will have a very rudimentary example of NetVR that contains these patents in a network visualization with a 3D gesture interaction that will allow certain nodes to be extruded from the plane of the visualization. The natural 3D gesture interaction is made possible by the Duke immersive Virtual Environment (DiVE). The DiVE is a 3m x 3m x 3m room where all six sides (walls, floor and ceiling) are rear-projected with stereoscopic images. Gesture and walking interactions are made possible through the use of tracking technology which is integrated into the graphical rendering system. The March 2007 version of NetVR will provide a simplistic node-link representation of the fixed set of patent data. An operator will be able to "grab-and-drag" specific nodes to highlight relationships within the dataset. It will also provide a "details-on-demand" window that shows the patent information when a single node is selected.

Through this Common Fund grant, the NetVR toolkit will be expanded to access the patent data through a web interface, allowing a scholar to dynamically update and filter the data being displayed. NetVR will incorporate clustering software for automatic node placement based on relationship information, as well as a geospatial embedding of the data using Google Earth. The data driving the NetVR visualization will be expanded to include citation data. Finally, the ability to save any specific configuration of a network will be incorporated, allowing a scholar to produce an image or movie appropriate for a publication or talk.

Goal 3: Empirically validate the visualization-patent framework by using it to examine the biomedical industries of North Carolina and study the role of institutions (including institutions of higher education) in fostering or impeding innovation.

For this research infrastructure to be usable for further research into innovation, it will need to be further developed and documented. John Madden has been one of the leading voices for improved information analysis within Duke University Medical Center, a particularly pressing concern given the creation of the new Translation Center. He has worked extensively with SNOMED using the same sort of "deep" semantic analysis that Tim Lenoir studies for taxonomizing and translating biomedical research. In order to discover networks of innovation within biomedical research, a combination of first-hand scientific knowledge of biomedical research and practice with a historical sensibility is needed. Therefore, the interdisciplinary combination of a scientist like Madden with a historian of biomedicine like Tim Lenoir is essential. Together, they will use the combination of visualization and patent database to explore the data, discovering the emergence of new platforms and technologies in biomedicine. Cathy Davidson will provide "meta-level" reflection, analysis, and administration of the research. Davidson will ensure that the essence of the research findings can be translated across disciplinary boundaries and that the process of collaboration is thoroughly documented.

Goal 4: Host a research symposium on the visualization of patents targeted to the Duke community in Spring 2008.

The network data itself would be demonstrated after the project is completed at some point in Spring 2008 by a well-advertised and catered public research symposium in CIEMAS for the entire Duke community. Tim Lenoir and John Madden will present the importance of visualization to the work of mapping these networks of innovation in patent data. Finally, Cathy Davidson will give a presentation on the role of technology in interdisciplinary scholarship, using the visualization work as her leading example, and translating the findings of the research within the wider context of interdisciplinary research at Duke. Duke attendees will be encouraged to use this infrastructure created for this research project for their own research purposes. In order to give

[1] See visualcomplexity.org for examples of state-of-the-art 2D network visualizations

participants a hands-on experience for the research, Brady and Madden will offer a guided tour of the visualized patent data inside the DiVE itself. A research paper by a combination of Lenoir,Madden, Davidson, and Brady will be written to be presented at least one internationally-known conference, where a similar demonstration will be given.

3. Project Team

This proposal combines Tim Lenoir's current research mapping and analyzing innovation using patents with Rachael Brady's visualization facilities in order to let researchers discover and map these networks of innovation in an intuitive and visual manner. John Madden has extensive experience in developing semantic search capabilities for fields in biomedicine. Madden will work with Tim Lenoir in developing an ontology and classification schemes that will form the basis of a powerful semantic search capability for an integrated data mashup of social, co-author, paper-citation, patent, and funding networks for biotechnology in the Research Triangle. Lenoir and Madden will work closely with Kevin Webb and Harry Halpin (a researcher who helped develop the infrastructure for other semantic web and the first generation of patent visualization technologies) in developing the semantic search engine for extracting networks of interconnection that support the biotech knowledge economy in the Research Triangle. Cathy Davidson will study the collaboration itself in order to use it as a model for her current work with the McArthur Foundation. Davidson's work with the MacArthur Foundation and HASTAC is dedicated to proposing cross-disciplinary and cross-institutional models of collaborative research and to understanding the most effective ways that quantitative data can support qualitative and theoretical modes of explanation and vice versa. She will play a similar role in this project, including writing up the results of this collaboration as one of the models highlighted in her MacArthur project on "The Future of Learning Institutions in a Digital Age."

4. Relationship to University Priorities

This project ties in with Duke's major new Visual Studies initiative by using sophisticated visualization tools as a means to better understand complex (and sometimes seemingly contradictory) data sets that will allow historians new insights into relationships between individuals, institutions, legal processes (in this case, patents), and innovation [6, 7]. In short, a new visualization apparatus will, ideally, yield a new kind of history of ideas by making visible (literally) relationships that we have not been able to see before. NetVR is meant not only to visualize patent networks per se, but any sort of networks. For example, one could imagine it being used to analyze networks of social relationships in Chinese villages, and to enable this it takes as its input a generic format for networks that is usable across different machines and disciplines. It could also be used as a tool for imagining intellectual relationships across disciplinary, departmental, and institutional boundaries across many universities. Also, an innovative algorithm to do a connection analysis of the network will be created, that allows multiple parts of the network to be selected and then the connections between them to "brighten," making it easy to see connections between patents, academic papers, people, and institutions. This visualization infrastructure, once fully documented and developed, will be an immense boon to further interdisciplinary scholarship at Duke. The results of this research will be presented at the symposium to the wider Duke community and at international academic venues, generating even more research from Duke University and international recognition.

Appendix: References

1. Börner, K., J. Maru, and R. Goldstone, *The Simultaneous Evolution of Author and Paper Networks.* Proceedings of the National Academy of Sciences USA, 2004. 101(Suppl_1): p. 5266-5273.
2. Börner, K., C. Chen, and K. Boyack, *Visualizing Knowledge Domains*, in *Annual Review of Information Science & Technology*, B. Cronin, Editor. 2003, Information Today, Inc./American Society for Information Science and Technology: Medford, NJ. p. 179-255.
3. Boyack, K.W. and K.B. Börner, *Indicator-Assisted Evaluation and Funding of Research: Visualizing the Influence of Grants on the Number and Citation Counts of Research Papers.* Journal of the American Society for Information Science and Technology, 2003. 54(5): p. 447-461.
4. Card, S., J. Mackinlay, and B. Shneiderman, eds. *Readings in Information Visualization: Using Vision to Think*. 1999, Morgan Kaufmann.
5. Chen, C., *Mapping Scientific Frontiers*. 2002, London: Springer-Verlag.
6. Dikovitskaya, Margaret (2005 (cloth), 2006 (paperback)). *Visual Culture: The Study of the Visual after the Cultural Turn*, 1st ed., Cambridge, Ma: The MIT Press.
7. Fuery, Kelli & Patrick Fuery (2003). *Visual Culture and Critical Theory*, 1st ed., London: Arnold Publisher.
8. Havre, S., B. Hetzler, and L. Nowell. *ThemeRiver: Visualizing Theme Changes over Time*. in *IEEE Symposium on Information Visualization, InfoVis 2000*. 2000: IEEE Press.
9. Lenoir, Timothy, et al., Inventing the Entrepreneurial Region: Stanford and the Co-Evolution of Silicon Valley, Stanford; Stanford University Press, 2007 (in press).
10. Mane, K. and K. Börner, *Mapping Topics and Topic Bursts in PNAS.* Proceedings of the National Academy of Sciences of the United States of America, Vol. 101 (2004): 5287-5290.
11. Morris, S., DeYong C, Wu Z, Salman S, Yemenu D., *DIVA: a visualization system for exploring document databases for technology forecasting.* Computers & Industrial Engineering, 2002. 1(43): p. 841–862.
12. Newman ME. *The structure of scientific collaboration networks.* Proc Natl Acad Sci U S A. 2001 Jan 16; 98(2):404-9.
13. Newman ME. *Coauthorship networks and patterns of scientific collaboration.* Proc Natl Acad Sci U S A. 2004 Apr 6; 101 Suppl 1:5200-5.
14. Newman ME. *Modularity and community structure in networks*. Proc Natl Acad Sci U S A. 2006; 103(23):8577-82.
15. Rinia, E.J., et al., *Measuring Knowledge Transfer between Fields of Science.* Scientometrics, 2002. 54(3): p. 347-362.
16. Shen, Z., Ogawa, M., Teoh, S. T., and Ma, K. 2006. BiblioViz: a system for visualizing bibliography information. In *Proceedings of the 2006 Asia-Pacific Symposium on information Visualisation - Volume 60* (Tokyo, Japan). K. Misue, K. Sugiyama, and J. Tanaka, Eds. ACM International Conference Proceeding Series, vol. 164. Australian Computer Society, Darlinghurst, Australia, 93-102.
17. Skupin, A. *From Metaphor to Method: Cartographic Perspectives on Information Visualization*. in *Proceedings of InfoVis 2000*. 2000. Salt Lake City, UT: IEEE Computer Society.
18. White, H.D. and K.W. McCain, *Visualization of literatures.* Annual Review of Information Science and Technology, 1997. 32: p. 99-168.
19. White, H.D. and K.W. McCain, *Visualizing a discipline: An author co-citation analysis of information science, 1972-1995.* Journal of the American Society for Information Science, 1998. 49(4): p. 327-356.