Spatio-Temporal Information Production and Consumption of Major U.S. Research Institutions[1]

Katy Börner and Shashikant Penumarthy

katy@indiana.edu, sprao@indiana.edu
School of Library and Information Science, Indiana University
10th Street & Jordan Avenue, Bloomington, IN 47405, USA

Abstract
This paper reports the results of a large scale data analysis that aims to identify the information production and consumption among top research institutions in the United States. A 20-year publication data set was analyzed to identify the 500 most cited research institutions and spatio-temporal changes in their inter-citation patterns. A novel approach to analyzing the dual role of institutions as information producers and consumers and to study the diffusion of information among them is introduced. A geographic visualization metaphor is used to visually depict the production and consumption of knowledge. Surprisingly, the introduction of the Internet does not seem to affect the distance over which information diffuses as manifested by citation links. The citation linkages between institutions fall off with the distance between them, and there is a strong linear relationship between the log of the citation counts and the log of the distance. The paper concludes with a discussion of these results and an outlook for future work.

Introduction

Does space still matter in the Internet age? Does one still have to study and work at major research institutions in order to have access to high quality data and expertise and to produce high quality research?

To answer these questions, an interdisciplinary publication data set covering the years from 1982-2001 was analyzed to identify the 500 most cited research institutions in the United States and spatial changes in their inter-citation patterns. Advanced data analysis and visualization techniques were applied to determine information sources and sinks and the diffusion patterns among them.

The results of our analysis are surprising in that the increasing usage of the Internet does not lead to more global citation patterns. In particular, the distance over which information diffuses as manifested by citation links does not increase over time.

The remainder of the paper is organized as follows: Section 2 reviews related work and contrasts it with our approach; Section 3 describes the data set used in this analysis and how it was processed; Visualizations of the data set are presented in section 4; Section 5 concludes the paper with a discussion of results and future work.

Related Work and Our Approach

The diffusion of tangible objects (people, goods, etc.) but also of intangible objects (ideas, activity levels, etc) has been studied in diverse fields of science including physics, e.g., heat diffusion; robotics, e.g., communication among mobile robots (Arai, Yoshida et al. 1993); social network analysis (Granovetter 1973; 2002); bibliometrics/scientometrics/webometrics (Katz 1994; Thelwall 2002), geography, e.g., migration studies (Ravenstein 1885; Thornwaite 1934; Tobler 1995); and biology, e.g., neuronal migration in the nervous system (Thurner, Wick et al. 2002).

Other studies have attempted to judge the research vitality or quality of research conducted at specific research institutions. Diverse activity, impact, and linkage measures exist and can be applied to quantify the research contribution of institutions (Narin, Olivastro et al. 1994). However, very few citation studies have attempted to analyze the geographical concentration of highly cited authors, institutions, countries. Batty's (2003) work is an exception and it nicely shows that the distribution of

[1] This research is supported by Career grant no. IIS-0238261 from the National Science Foundation awarded to the first author. We gratefully acknowledge support from the Center for the Study of Institutions, Population, and Environmental Change at Indiana University through National Science Foundation grant BCS-0215738.

Proceedings of ISSI 2005, pages 635–641.
Edited by P. Ingwersen and B. Larsen.

citation counts is highly skewed, with most citations being associated with a few individuals working at a small number of institutions in an even smaller number of places and countries.

Here, we are interested to study the diffusion of scholarly knowledge. We assume that scholarly knowledge diffuses via co-authorships, the physical movement of authors through geographical space and the production (writing) and consumption (citing) of papers, among others. Unfortunately, the identification of unique author names is unresolved. Similarly, proper contribution of an author to his or her institution is often impossible due to the quality of available publication data.

Our work goes beyond existing research in that we do not only examine the citation counts for each institution but attempt to (1) identify geographically and statistically significant instances of institutions that act as major information sources, (2) correlate their behavior as *information sources* (number of citations their papers receive), *information sinks* (number of references to papers produced at other institutions), and *self-consumers* (number of self citations), (3) use direct citation linkage to identify their interrelation based on the amount of directly exchanged information, and (4) analyze and visualize the importance of proximity in geographic space for information exchange.

Subsequently, we formalize each institution as a node that acts as both: a source (or producer) of information as well as an information sink (or consumer). Arrows among institutions denote the flow of information. If a paper was published at institution A and is cited by a paper that is published at institution B, then there will be an arrow going from A to B. The more papers produced at A are cited by B, the higher the volume of information flow. Hence, the normalized out-degree of a node can be used to characterize the role of an institution as an information source. The normalized in-degree of a node describes the role of an institution as an information sink. Links which lead from an institution to itself correspond to self-citations. Note that this formalization could also be applied to authors, countries, etc.

Data Set and Data Analysis

The complete set of papers published in the Proceedings of the National Academy of Sciences (PNAS) in the years from 1982-2001 was analyzed to determine knowledge diffusion pathways among major institutions as manifested in paper citation linkages among the papers. The data set contains 47,073 papers published by 18,994 unique authors, who work at 2,822 institutions. Institutions comprise academic institutions, research labs and corporate entities. To be credited with an article, a given institution had to be the site of the first author listed on the paper. The paper most highly cited by papers within the set received 612 citations.

Given our interest in exploring the importance of spatial proximity for the diffusion of information within U.S., we decided to analyze information diffusion patterns among major institutions, the spatial position of which is uniquely and persistently identified by their zip code and corresponding longitude and latitude coordinates. By ‘major institutions’, we refer to institutions that have acquired the highest total number of citations for their papers.

An initial data cleaning step was performed to remove suffixes such as INC, MED. These suffixes serve to indicate whether the entity in question is a corporate entity, a research lab or an academic institution. However, these suffixes are not consistent with respect to spacing between the name of the institution and the suffix, leading to string matching problems. Removing these suffixes helps to create uniformity of institution names in the data set.

Next, we had to decide what institutions should be merged. For example, an institution such as Indiana University has several campuses. Collapsing all these campuses into one entity causes valuable geographic information to be lost, since the campuses might be far apart. However, separating out each campus individually can result in extremely cluttered data. Another significant issue that arises out of separating different campuses of the same university is the distribution of the number of citations among those campuses. For example, Indiana University as a single entity might qualify to be in the top 500 most highly cited institution list, but when the campuses are split, none of the individual campuses might have the requisite number of citations to make it into this list.

The zip code was used to preserve information about where two institutions with the same name, but with differing geographic locations, are located. The United States zip code assigns postal codes based on the position of a certain geographic location in a hierarchy of geographic significance based on area. Hence, in the 5-digit zip code, the first digit indicates which region of the U.S. the location belongs to such as northeast, southwest, etc. The next two digits indicate state and county

information. The final two digits serve to distinguish finer boundaries such as towns and cities within a county. A unique ID was created for each institution by concatenating the (abbreviated) name of the institution with its zip code. As this system is unique to the United States, non-U.S. institutions, such as University of Tokyo (1,797 citations), despite producing highly cited publications, were excluded from the analysis presented in this paper.

We then proceeded to determine the level of geographic resolution that is significant for answering our question. Given that universities typically do not have two major campuses in one county we decided to use the county as our smallest unit. Hence, for each institution, all its campuses or instances that lay within the same county were collapsed into one entity. In zip code terms, this meant merging all instances of an institution whose zip codes differed only in the last two digits. The newly created identity of the institution consisted of a concatenation of the (abbreviated) name with the smallest zip code within that county. For example, INDIANA UNIV47401 and INDIANA UNIV47405 were collapsed into INDIANA UNIV47401. Collapsing universities in this manner provides a good compromise between maintaining geographic identity and statistical significance.

Subsequently, the top 500 most highly cited institutions were identified. The top 500 institutions produced 30,572 (64.95%) of all papers and received 195,889 (51.83%) of a total of 377,935 citations. A graph showing the number of listed references, received citations, and self citations over the alphabetically sorted list of institutions is given in Figure 1. An offset was applied to citation counts to improve readability.

Exactly five institutions produced papers that attracted more than 4,000 citations. Harvard (HARVARD UNIV02114) leads with 13,763 citations. MIT (MIT02139) follows with 5,261. Johns Hopkins University (JOHNS HOPKINS UNIV21201) has 4,848. STANFORD UNIV94302 accumulated 4,546 and UNIV CALIF SAN FRANCISCO94103 got 4,471.

For each institution we determined the ratio of the number of citations received by this institution divided by the sum of received citations and references made, multiplied by 100. Interestingly, there are 131 institutions with a value between 0-40% acting mostly as information producers. 71 of the institutions have a value between 60-100% and act mostly as information consumers – they reference a large number of papers but the number of citations they receive is comparably low.

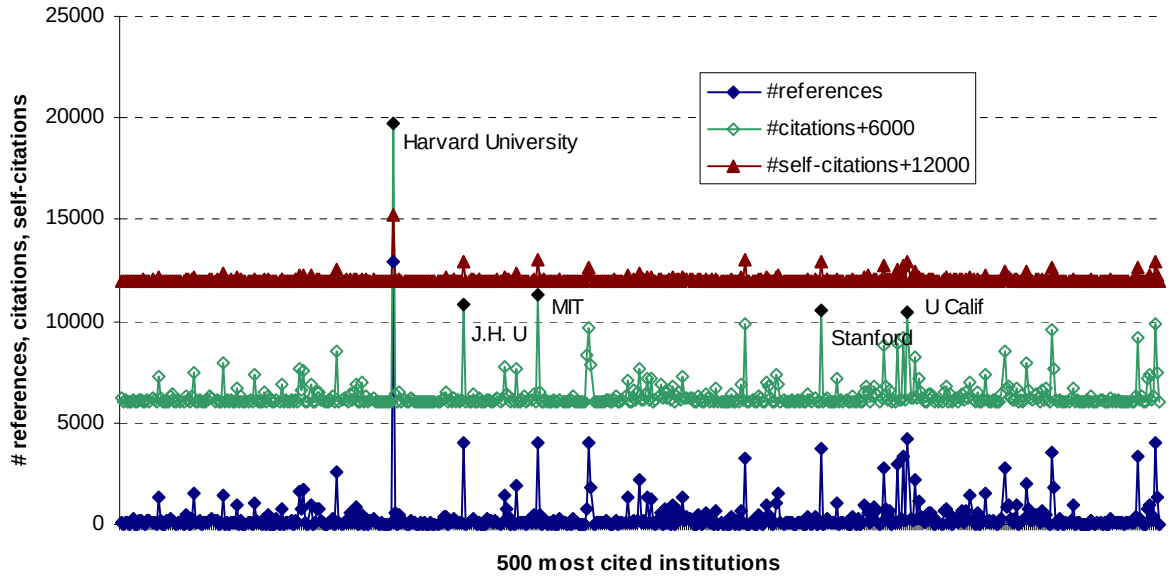

Figure 11: Number of listed references, received citations, and self-citations

Next, we examined the very unsymmetrical direct citation linkage patterns among the top 500 institutions. A visual depiction of the result is given in Figure 2. The high peak values in the diagonal reflect the high amount of self-citations for all institutions. The medium peak horizontal and vertical lines denote references from and citations to papers written at Harvard University.

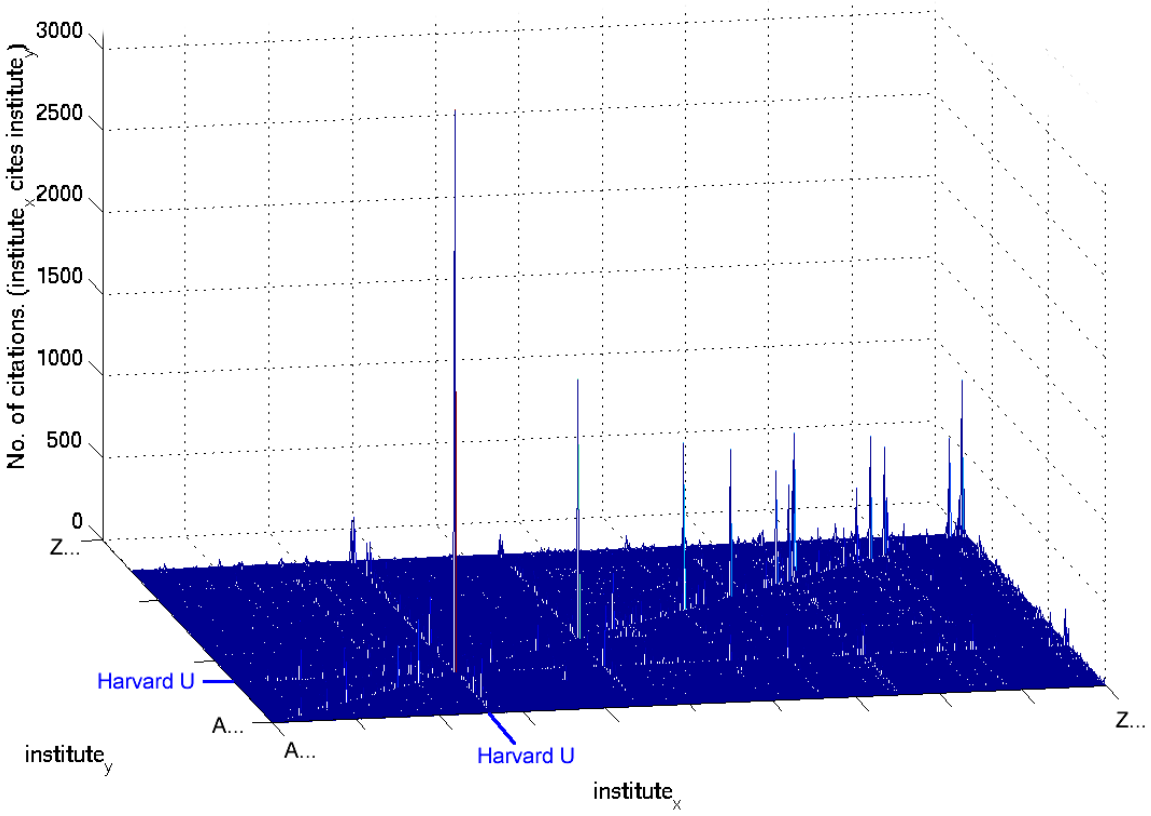

Figure 12: Institution cross citation matrix for the top 500 most cited institutions

Geographic Visualizations

The ArcGIS program from ESRI's geographic information system (GIS) was applied to show the geographic distribution of the top 500 institutions in geographic space. ESRI's geocoding service translates U.S. zip codes into latitude and longitude information using the Albers equal area projection, thus preserving the earth's surface area.

While the GIS is highly interactive, allowing users to get an overview of the data, zoom into a subset or subarea and to get details on demand (Shneiderman 1996), the visualizations presented in this paper are static snapshots of the system interface. However, they were optimized to show complex citation patterns despite their static and two-dimensional appearance.

Figure 3 shows a map of U.S with states color coded based on the population size in the year 2000. Lighter shades of green represent lower population. Overlaid are the top 500 institutions. Each institution is represented by 'citation stick'. The color of the stick corresponds to the number of citations that institutions received from other institutions in the 500 item data set over the 20-year time span, see legend in the right lower part of Figure 3.

The stick height is a function of the normalized number of citations received for a certain institution in relation to the maximum number of citations that any institution received:

$$height = \left(\sin\left(\frac{\#citations}{\max\#citations} \right) + 1 \right) * k .$$

The utilization of sin guaranties that small differences between institutions with low citation counts are visible and that the huge differences among the institutions with high citation counts are less distorting. k is a scaling factor.

Harvard University clearly has the highest number of citations and hence max# citations equals 13,763 (excluding self citations). It is followed by MIT and Johns Hopkins with 5,261 and 4,848 citations respectively. This conforms with work by Adams (1998) who showed that Harvard tops in

scientific impact by not only churning out more papers than any other university between 1993 and 1997, but also by producing work that was rated as having higher scientific impact across the board.

Figure 3: Geographic location and number of received citations for the top 500 institutions

Using Tobler's (1995) analogy of flow of energy in a vector potential field, highly cited institutions exhibit a high pressure for the diffusion of information whereas other institutions are mostly importing information and hence act as information sinks.

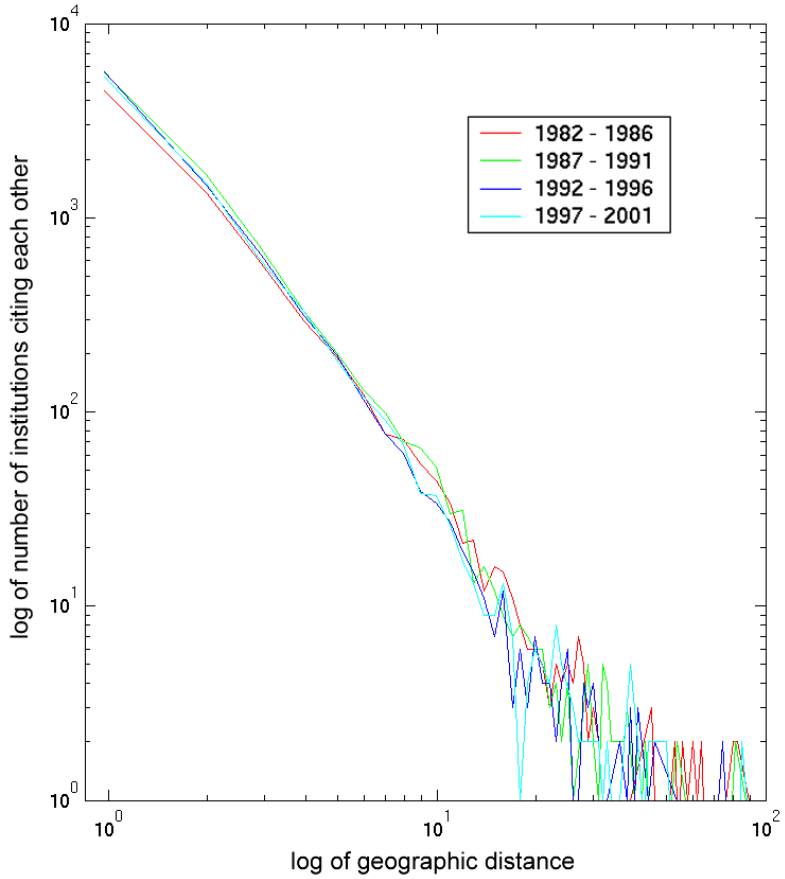

Figure 4: Log-log plot showing the variation of the number of institutions that cite each other over geographic distance among them for each of the four time slices. The distance was calculated by applying the Euclidean form formulae to xy coordinates obtained using the Albers projection. 1.5 units of geographic distance equal approximately 100 miles.

In addition, we were interested to see if there were major changes in the distributions of the number of institutions that cite each other across a certain geographic distance over time. To

investigate this question, we merged the geographic distance and number of received citations matrices into a list for each of the four time slices, and then sorted the lists by geographic distance. We then binned the geographic distance (in our scale of geographic distance each bin takes a range of 0.0926) and determined the number of institutions citing each other within each range of geographic distance. The resulting log-log graph is given in Figure 4.

The best fitting power law exponent for the years 82-86 is 1.94 and the power law accounts for 91.5% of the variance. Values for years 87-91 are 2.11 (93.5%), years 92-96 are 2.01 (90.8%), years 97-01 are 2.01 (90.7%). This result is rather surprising. As time progresses, and the amount of produced papers increases, space seems to matter more. Authors are more likely to cite papers generated by authors at close-by institutions.

One possible explanation could be that when flooded by information, the social component in citation (Wellman, White et al. 2004) – the importance of having interpersonal as well as intellectual ties – becomes more important. Obviously, a trip from Boston to San Francisco is less prudent than one to Washington D.C.

Discussion

The presented analysis provides a novel approach to analyzing the dual role of institutions as information producers and consumers and to study the diffusion of information among them. We hope this paper inspires similar studies on, e.g., the dual role of authors as information producers and consumers or the diffusion of information among companies via publication and patent citations, email exchanges, etc.

The results are rather counterintuitive and need to be examined in more detail before final conclusions can be made. It will be interesting to study why the introduction of the Internet does not lead to a more global citation behavior. Reasons for local collaborations might comprise 'winner takes all' funding schemes, the demands of complex, large-scale instrumentation, and the need to gain experience, train researchers, and sponsor protégés, see also (Katz 1994), p. 32.

We believe advanced information analysis and visualization techniques will be critical to understand the dynamics of information diffusion. Of particular interest will be techniques that can visualize diffusion patterns among many different static or moving instances. A first attempt to visualize social diffusion patterns was made in (Börner and Penumarthy 2003). Our future work will address the analysis and visualization of diffusion patterns of tangible and intangible objects over space and time.

We are aware that this first analysis has a number of shortcomings due to the coverage and quality of the used data set. While the PNAS data set nicely represents major research results from diverse areas of science over a 20 year time span, it does not cover any specific discipline completely nor does it represent any authors' entire life work. In the PNAS data set (and most other publication data sets) there is no means to attribute a certain percentage of a paper to each co-author (and his/her institution). Non-U.S. institutions had to be excluded from this analysis as no information about their longitude/latitude information was available to us. Obviously, the number of co-authorships or co-PI-ships, co-citations of papers, and co-occurrence of words in papers are additional valid indicators for information diffusion among institutions. Again, these issues open a number of interesting avenues for future research.

Acknowledgements

We would like to thank Tom Evans for making Zip code data available to us and Nathan Eaton for providing us with the complete ESRI U.S. data set.

References

Adams, A. (1998). "Citation analysis: Harvard tops in scientific impact." Science **281**(5385): 1936-1936.

Arai, T., E. Yoshida and J. Ota (1993). Information diffusion by local communication of multiple mobile robots. IEEE Conference on Systems, Man and Cybernetics.

Bankes, S. C. (2002). "Agent-based modeling: A revolution?" Proc. Natl. Acad. Sci. USA **99**(Suppl. 3): 7199-7200.

Batty, M. (2003). "The geography of scientific citation." Environ Plan A **35**: 761-765.

Börner, K. and S. Penumarthy (2003). "Social Diffusion Patterns in Three-Dimensional Virtual Worlds." Information Visualization **2**(3): 182-198.

Granovetter, M. (1973). "Strength of Weak Ties." American Journal of Sociology **78**: 1360--1380.
Katz, J. S. (1994). "Geographical proximity in scientific collaboration." Scientometrics **31**(1): 31-43.
Narin, F., D. Olivastro and K. A. Stevens (1994). "Bibliometrics? Theory, Practice and Problem." Evaluation Review **18**(1): 65-76.
Ravenstein, E. G. (1885). "The Laws of Migration." Journal of the Statistical Society **48**(2): 167-235.
Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualization. Proceedings of the Workshop on Visual Language, Boulder, CO, IEEE.
Thelwall, M. (2002). "Evidence for the existence of geographic trends in university web site interlinking." Journal of Documentation **58**(5): 563-574.
Thornwaite, C. W. (1934). Internal Migration in the U.S. Philadelphia, University of Pennsylvania Press.
Thurner, S., N. Wick, R. Hanel, R. Sedivy and L. A. Huber (2002). "Anomalous diffusion on dynamical networks: A model for interacting cell migration." Physica A.
Tobler, W. (1995). "Migration: Ravenstein, Thornthwaite, and Beyond." Urban Geography **16**(4): 327-343.
Wellman, B., H. D. White and N. Nazer (2004). "Does Citation Reflect Social Structure? Longitudinal Evidence from the 'Globenet' Interdisciplinary Reserach Group." Journal of the American Society for Information Science and Technology **55**(2): 111-126.