Börner, Katy, Penumarthy, Shashikant, Meiss, Mark and Ke, Weimao.
(2006) Mapping the Diffusion of Information Among Major U.S. Research Institutions.
Scientometrics. Vol. 68(3), pp. 415-426.

Mapping the Diffusion of Information Among Major U.S. Research Institutions

Katy Börner
School of Library and Information Science, Indiana University, 10th Street & Jordan Avenue, Bloomington, IN 47405, USA. E-mail: katy@indiana.edu
Shashikant Penumarthy
School of Library and Information Science, Indiana University, Bloomington, IN 47405. E-mail: sprao@indiana.edu
Mark Meiss
School of Informatics, Indiana University, Bloomington, IN 47405. E-mail: mmeiss@indiana.edu
Weimao Ke
School of Library and Information Science, Indiana University, Bloomington, IN 47405. E-mail: wke@indiana.edu

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. The highest producers and their consumers as well as the highest consumers and their producers are identified and mapped. 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.

This is a revised and extended version of a paper that was originally presented at the 10th International Conference of the International Society for Scientometrics and Informetrics in Stockholm, July 2005 [1].

Introduction

Does space still matter in the Internet age? Does one still have to study and work at a major research institution in order to have access to high quality data and expertise, to produce high quality research, and to diffuse results effectively?

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 seem to 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 analysis results 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 [2]; social network analysis [3, 4]; bibliometrics/scientometrics/webometrics [5, 6], geography, e.g., migration studies [7-9]; and biology, e.g., neuronal migration in the nervous system [10].

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 [11]. However, very few citation studies have attempted to analyze the geographical concentration of highly cited authors, institutions, countries. Batty's [12] work is an exception and it nicely shows that the distribution of 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 assignment 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 *producers* or *information sources* representing the number of citations their papers receive, *consumers* or *information sinks* based on the number of citations they make to papers produced at other institutions, and *self-consumers* reflecting the 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, institutions, 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 a high 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.

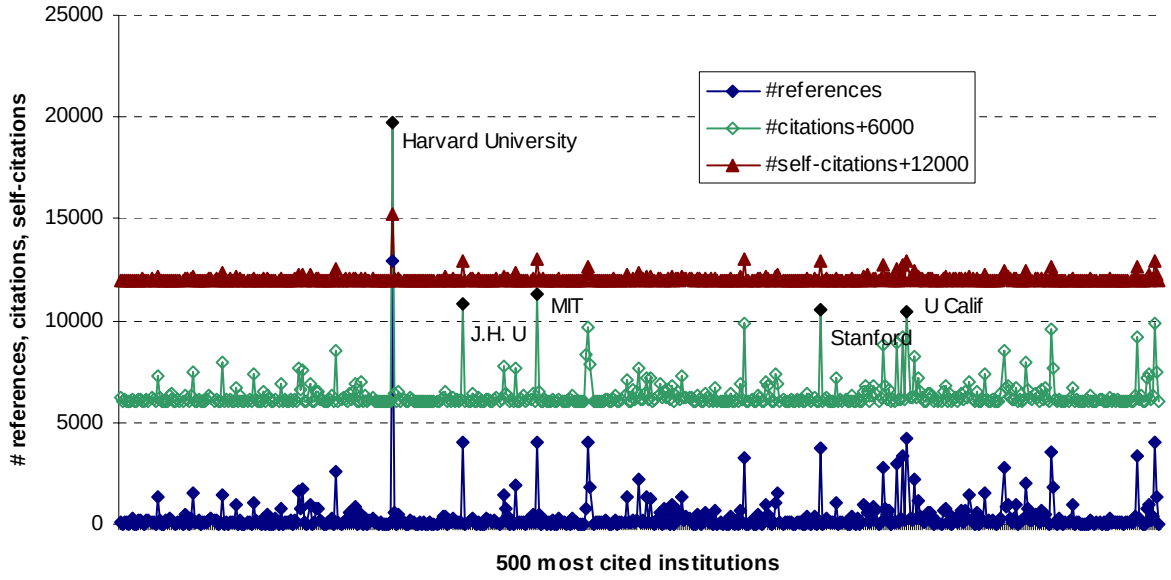

Figure 1: Number of listed references, received citations, and self-citations

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.

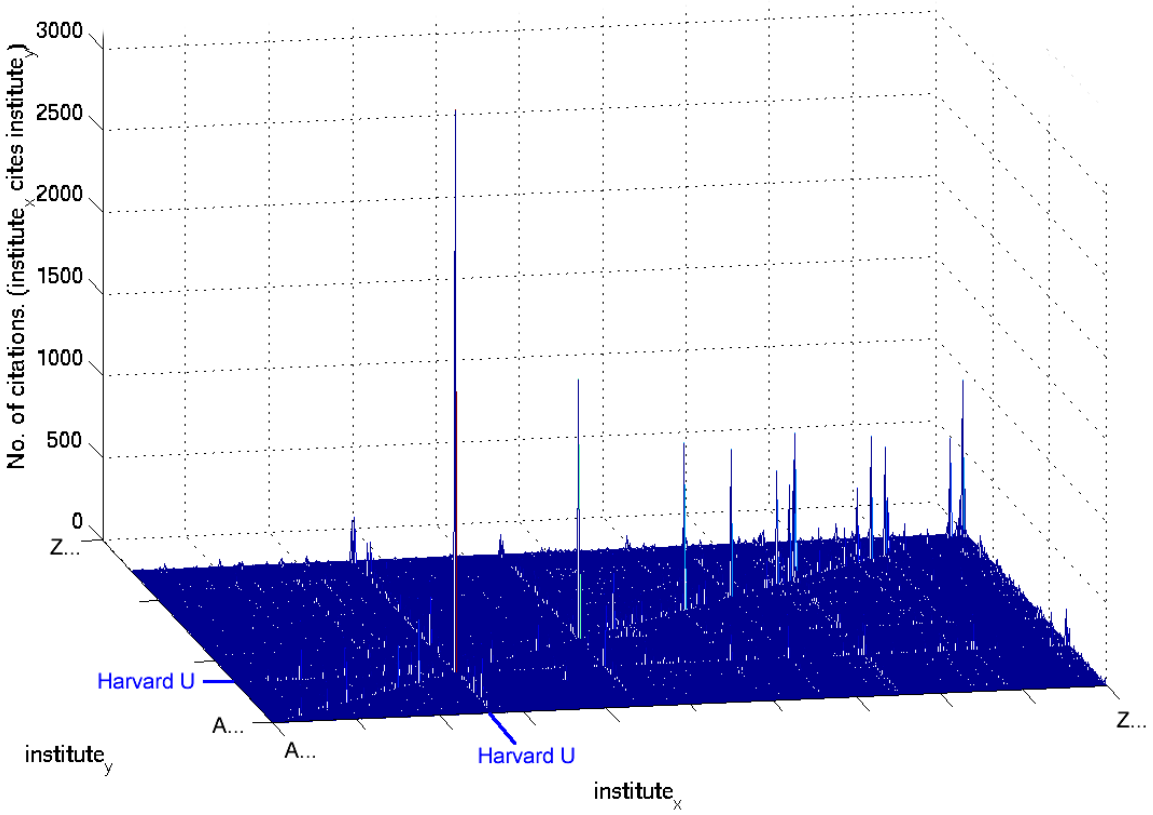

Figure 2: Institution cross citation matrix for the top 500 most cited institutions

Finally, we determined the ratio of the number of citations received by an 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. Using Tobler's [9] analogy of flow of energy in a vector potential field, highly cited institutions can be said to exhibit a high pressure for the diffusion of information whereas other institutions are mostly importing information and hence act as information sinks.

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 [13], 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 a '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.

Exactly five institutions produced papers that attracted more than 5,000 citations (excluding self citations) and are listed in Table 1. Harvard clearly leads with 16,531 citations (excluding self citations). This conforms with work by Adams [14] 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

Geospatial Distribution of Citations

To better understand the exchange of information among the different institutions, we determined the top five producers and their top ten consumers, see Table 1, as well as the top five consumers and their top 10 producers, see Table 2. The geospatial flow among these top producers and consumers is depicted in Figures 4 and 5.

Table 1: Top five producers, their total number of received citations (excluding self citations), and their top ten consumers

Producers, i.e., cited institutions	# citations received	Top ten consumers, i.e., institutions that cite institution listed in first column ordered by decreasing number of citations made
Harvard U	16,531	MIT, Brigham & Womens Hosp, Massachusetts Gen Hosp, Washington U, Yale U, Johns Hopkins U, NCI, U Washington, U Calif SF, Stanford U
MIT	7,033	Harvard U, YALE U, Whitehead Inst Biomed RES, U Calif SF, U Calif LA, U Washington, NCI, U Calif SD, Washington U, Massachusetts Gen Hosp
Stanford U	5,965	Harvard U, U Calif Berkeley, MIT, Yale U, U CALIF SF, U Washington, Washington U, U Calif SD, NCI, Johns Hopkins

		U
U Calif SF	5,779	Washington U, Stanford U, MIT, U Calif Berkeley, U Calif LA, U Calif SD, U Washington, Johns Hopkins U, U Texas
Johns Hopkins U	5,755	Harvard U, MIT, NCI, U Calif SF, Yale U, U Calif SD, Washington U, Stanford U, U Calif LA, U Calif Berkeley

Table 2: Top five consumers, their total number of citations made (excluding self citations), and their top ten producers

Consumers, i.e., citing institutions	# citations made	Top ten producers, i.e., institutions that are cited by institution listed in first column ordered by decreasing number of citations received.
Harvard U	13,552	MIT, Massachusetts Gen Hosp, Brigham & Womens Hosp, Johns Hopkins U, Stanford U, U Calif San Francisco, Yale U, Rockefeller U, U Washington, Washington U
U Calif SF	4,682	Harvard U, MIT, Stanford U, Johns Hopkins U, U Washington, Washington U, U Calif Berkeley, U Texas, U Calif SD, U Calif LA
MIT	4,655	Harvard U, Whitehead Inst Biomed Res, Johns Hopkins U, Stanford U, U Calif SF, Yale U, Rockefeller U, U Calif LA, Massachusetts Gen Hosp, U Calif Berkeley
NCI (zip: 20814)	4,519	Harvard U, NCI (zip: 20205), NCI (zip: 21701), MIT, Duke U, Johns Hopkins U, NIAID NICHHD, Stanford U, U Calif SF
Yale U	4,464	Harvard U, MIT, Stanford U, Rockefeller U, Johns Hopkins U, Washington U, U Calif SF, U Washington, NCI, Massachusetts Gen Hosp

Next, we were interested to visually depict the geospatial diffusion of information among the institutions listed in Tables 1 and 2. The Chizu system [15] was extended to render a geographic base map and to overlay institutions and their information exchange patterns.

Figure 4 shows a map of U.S. with states color coded based on the total number of citation counts received by their institutions (excluding self citations). Overlaid on the map are the intercitation patters among the top five producers and their top ten consumers. Each of the five producers is represented by a dot that has a different color, e.g., Harvard U is given in yellow. The area size of the colored dot corresponds to the number of citations received, see Table 1. Citations are represented by lines that interconnect producers and consumers. Lines are shaded from colored (source of information) to white (sink of information). An enlarged version of the top three East Coast producers is given as well.

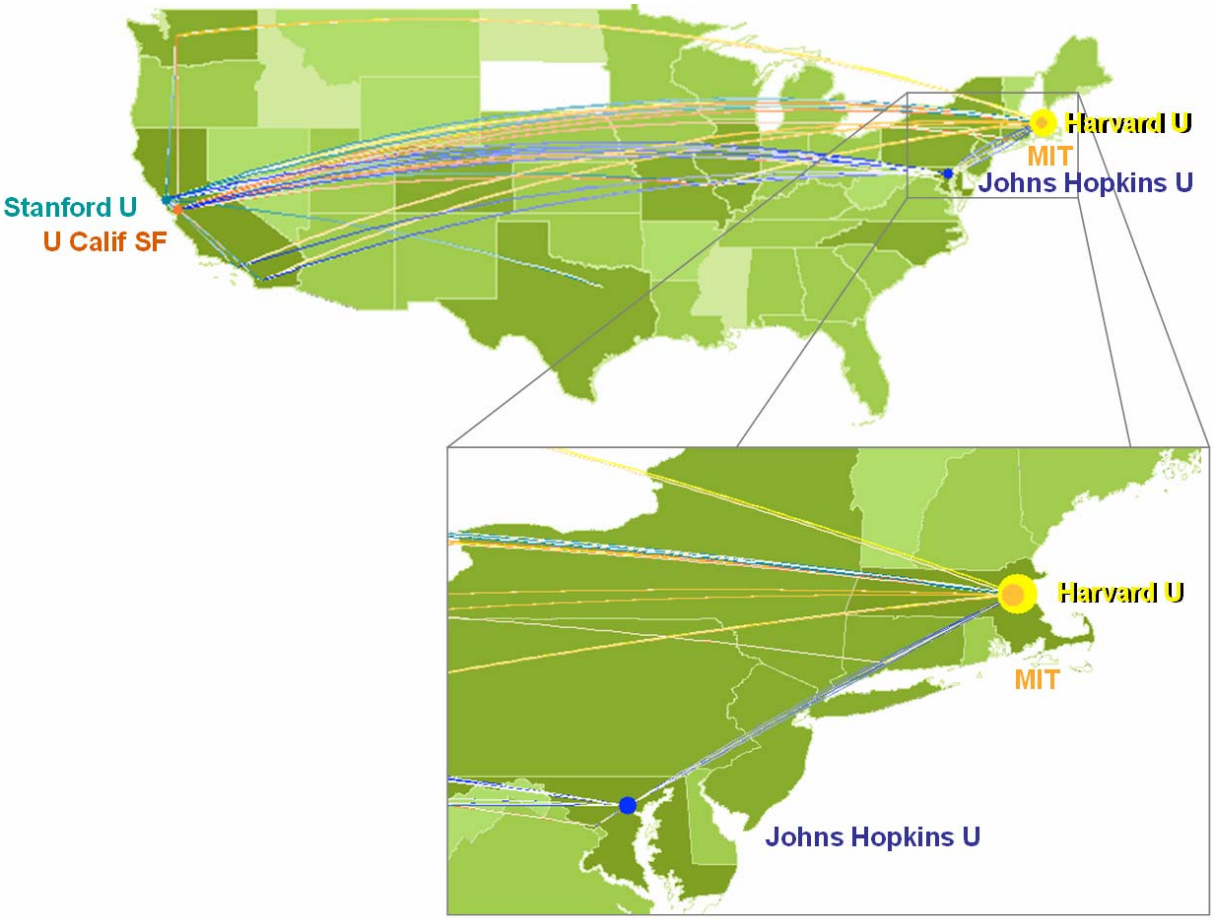

Figure 4: Geospatial information flow among the top five producers and their top ten consumers

Figure 5 shows that same map of U.S. but with states color coded based on the total number of citations made by the institutions in a state (excluding self citations). Overlaid are the intercitation patters among the top five consumers and their top ten producers. Each of the five consumers is represented in a different color, e.g., Harvard U is given in yellow. Like in Figure 4, the area size of the colored dot corresponds to the number of citations made, see Table 2. Citations are represented by lines that interconnect producers and consumers. Lines are shaded from black (source of information) to colored (sink of information).

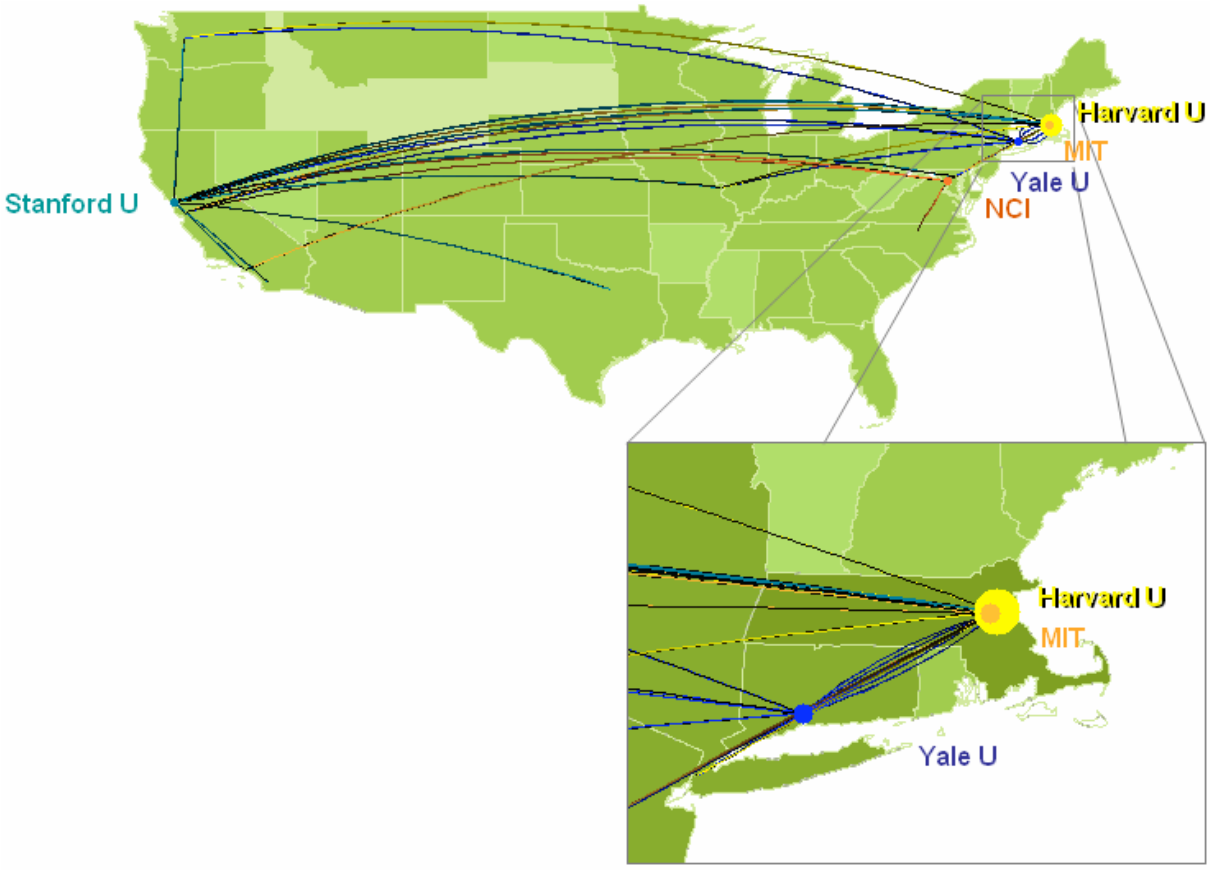

Figure 5: Geospatial information flow among the top five consumers and their top ten producers

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 divided the dataset in four time slices. Subsequently, we determined the geographic distance for all citations made in any of the four time periods. We then binned the geographic distance (in our scale of geographic distance each bin corresponds to about 6 miles) and determined the number of institutions citing each other within each range of geographic distance. The resulting log-log graph is given in Figure 6.