Hook, Peter A. (2007) Visualizing the Topic Space of the United States Supreme Court.
In Proceedings of the 11th International Conference on Scientometrics and Informetrics (ISSI 2007),
Madrid, Spain: June 25-27, pp. 387-396.

Visualizing the Topic Space of the United States Supreme Court[1]

Peter A. Hook

pahook@indiana.edu
Indiana University, School of Law, 211 South Indiana Ave., Bloomington, IN 47405 (United States)

Abstract
This article describes the creation of several domain maps based on the topic space of opinions issued by the United States Supreme Court. Topics assigned by West Publishing were harvested off of the Westlaw database and visualized using Principal Components Analysis (PCA), Multidimensional Scaling (MDS), and graph visualization software (Pajek). Peculiar topic adjacencies were noted and attributed to the unique nature of cases argued at the level of the United States Supreme Court. The work is contextualized throughout by the author's desire to create a rigorous base map on which to layer additional data for teaching purposes.

Keywords
Domain Maps; Visualizations; United States Supreme Court; Law; Political Science; Westlaw

Introduction

Background and Purpose
Scientometrics and bibliometrics owe a debt of gratitude to the legal research publishing industry in the United States. Frank Shepard's legal citator (Ogden, 1993) was part of the inspiration for Eugene Garlfield's *Science Citation Index* and subsequent products (Garfield, 1955 & 1979). This in turn was part of the inspiration for Page and Brin's PageRank algorithm—the foundation for Google (Hopkins, 2005; Battelle, 2005). Now, the tools of scientometrics may assist the legal research publishing industry to more optimally organize its materials. Legal information is itself exciting because it is one of the largest and most atomistically indexed bodies of information.

This research seeks to identify the topical adjacencies of subjects addressed in legal cases by the United States Supreme Court based on the co-occurrence of top level topics assigned by West Publishing (Thomson/West, 2006). It is in furtherance of the author's goal of creating a rigorous substrate map on which to layer over sixty years of Supreme Court topic data to be used for teaching purposes. In addition, the research is related to a growing body of work detailing and analyzing the network structure of legal opinions and their citation linkages (Chandler, 2005; Cross & Smith, In Press; Cross, Smith & Tomarchio, In Press; Fowler et. al., In Press; Smith, In Press), judicial and legislative co-voting networks (Fowler, 2006; Epstein et. al., 2005; Johnson et. al., 2005; Poole, 2005; Porter et. al., 2005; Sirovich, 2003; Brazill, 2002; Grofman, 2002; Martin & Quinn, 2002; Spaeth & Altfeld, 1985; Schubert, 1962 & 1963; Thurstone & Degan, 1951; Pritchett, 1941), and the move in legal academia toward quantitative empirical scholarship (George, 2006).

Maps of inherently non-spatial data that use a spatial substrate on which to layer additional information are common in information science (Hook & Börner, 2005). These maps employ the distance-similarity metaphor by which the viewer infers that items more proximate in space are more related than items further apart (Montello et al., 2003; Skupin and Fabrikant, 2003). The

[1] Full color images, the text of this paper, and additional appendixes are available at: http://ella.slis.indiana.edu/~pahook/index.html.

benefit of a substrate map is that it provides a common background from which changes may be readily perceived and is thus useful for pedagogy and illustrating changes over time.

Spatial layouts of inherently non-spatial data may be created in several ways. The first way is by the opinion of experts as to which topics are most similar and by laying out those topics by intuitive warrant or heuristics (*See* Bernal, 1939; Ellingham, 1948). The second way is by algorithmic comparisons of similarity and automated layouts using objective measures such as citation linkages or the co-occurrence of terms (Börner, Chen, & Boyack, 2002). Finally, a third method is a fusion approach which combines elements of each of the first two methods. For the most part, this paper employs multivariate statistical techniques that fall into the second category. These techniques are principal component analysis ("PCA") and multidimensional scaling ("MDS'). However, elements of the fusion approach were used when the author placed data elements into higher level categories based on his training in and experience of the United States legal system before employing the multivariate statistical techniques.

Methods, Materials, Procedures, and Equipment Used

Data Summary

The dataset used for this research consists of bibliographic information about all United States Supreme Court cases that have been issued West topics by West Publishing from the 1944 Term through the end of the 2004 Term (October 1944 through July 2005). The author harvested the data as an academic end user from the Westlaw database. The data contains information about 7,948 unique Supreme Court cases to which 19,789 topic assignments have been made. Of the 405 top level topics in the West taxonomy, 290 appear in opinions issued by the Supreme Court for this time period. All but one ("Reference"), co-occur with other topics resulting in 22,345 edges between cases sharing a similar topic. There are 3743 unique topic pairings.

About the Data

For over a hundred years, West Publishing has identified unique statements of law within court cases (Surrency, 1990). Human editors working at West assign these unique and legally controlling statements topic identifiers from its taxonomy of the law known as the West Topic and Key Number System (Doyle, 1992; Snyder, 1999; Thomson/West, 2006). Before the advent of online full-text searching, the West Topic and Key Number System was one of the only ways to research cases on a given issue. Now, the Topic and Key Number System is used primarily to augment free text searching and to convince a researcher that he or she has found all of the appropriate cases on a particular topic. The Westlaw Database, owned by Thomson/West Publishing, provides online access to United States Supreme Court opinions, numerous other cases, and additional legal material. It is a proprietary subscription database that includes both the actual language of court opinions plus editorial enhancements provided by West such as topic assignments from the West Topic and Key Number System.

Data Harvesting

The data was harvested off of the Westlaw database during March through April, 2004. As of March 18, 2004, there were 405 top level topics in the West Topic and Key Number System. A search as to each of the 405 topics was conducted by hand using the conventional end user interface. A typical search statement was: TO("2 Abatement and Revival"). The TO in this case means topic and the scope of the database at the time included all Supreme Court opinions from the 1944 term to date. The resultant list of cases for each of the 405 topic searches were placed into a spreadsheet along with the topic that caused the case to be returned by the database. Topic assignments were aggregated such that each case was listed with all of its topic assignments and did not appear more than once. Subsequent Supreme Court cases and their topic assignments

were added later. Annually, West makes changes to its taxonomy. In order for the dataset to include cases after the original March through April 2004 harvesting period, the author had to account for these changes. On several occasions, new topics were converted to their previous equivalents to bring the dataset current through the end of July 2005.

Additional Human Coding of the Data

Additionally, the author employed his legal training and knowledge of how concepts are taught in law school to make additional subject matter assignments to the 405 West topics: (1) Doctrinal – relevant to a specific subject taught in law school. (Constitutional Law, Administrative Law), (2) Factual – with unique factual circumstances relating to the topic but whose doctrinal elements are drawn from other topics (Aviation Law, Automobile Law), and (3) Procedural – capable of arising in almost any factual or doctrinal situation (Federal Courts, Federal Civil Procedure). For the doctrinal and procedural topics, the author also assigned categories to the topics based on in what course they are most likely to be covered in law school**.**

Data Manipulation and Visualization

The data was imported to the R statistical computing environment. Before applying the multivariate statistical visualization techniques, the data had to be put into matrix form. The data comprises a sparse matrix of 3743 unique topic pairings out of a theoretically possible 83,521 (289 x 289). The range of topic co-occurrence counts is 1 to 896 (with Constitutional Law and Federal Courts (896) being the most commonly co-occurring topics and Constitutional Law and Criminal Law (468) being the second most common). The mean topic co-occurrence count was only 5.97 and the median and mode were both 1. Both PCA and MDS were performed on the data. PCA was performed using Singular Value Decomposition (SVD). The resultant plots were useful to characterize the major dimensions in the variation in the data of topic co-occurrence. (See generally Paolillo and Wright, 2006). Additionally, the dataset was visualized in its network form using the network visualization and analysis tool, Pajek (Batagelj & Mrvar, 1998). In the parlance of network science, the nodes represented West Topics and the edges represented the co-occurrence of those topics in Supreme Court cases.

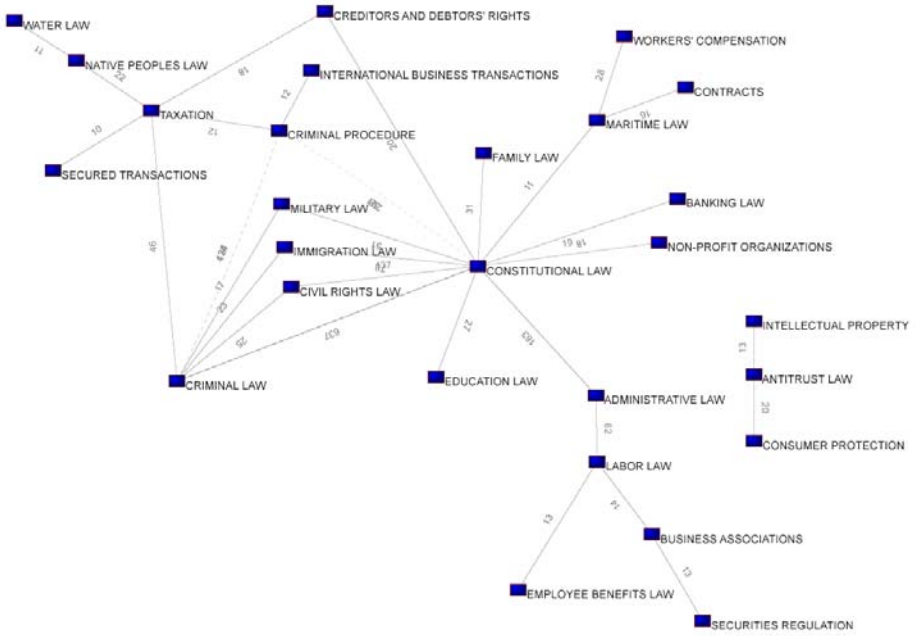


Figure 1: West Topic Space of the United States Supreme Court—Network Layout

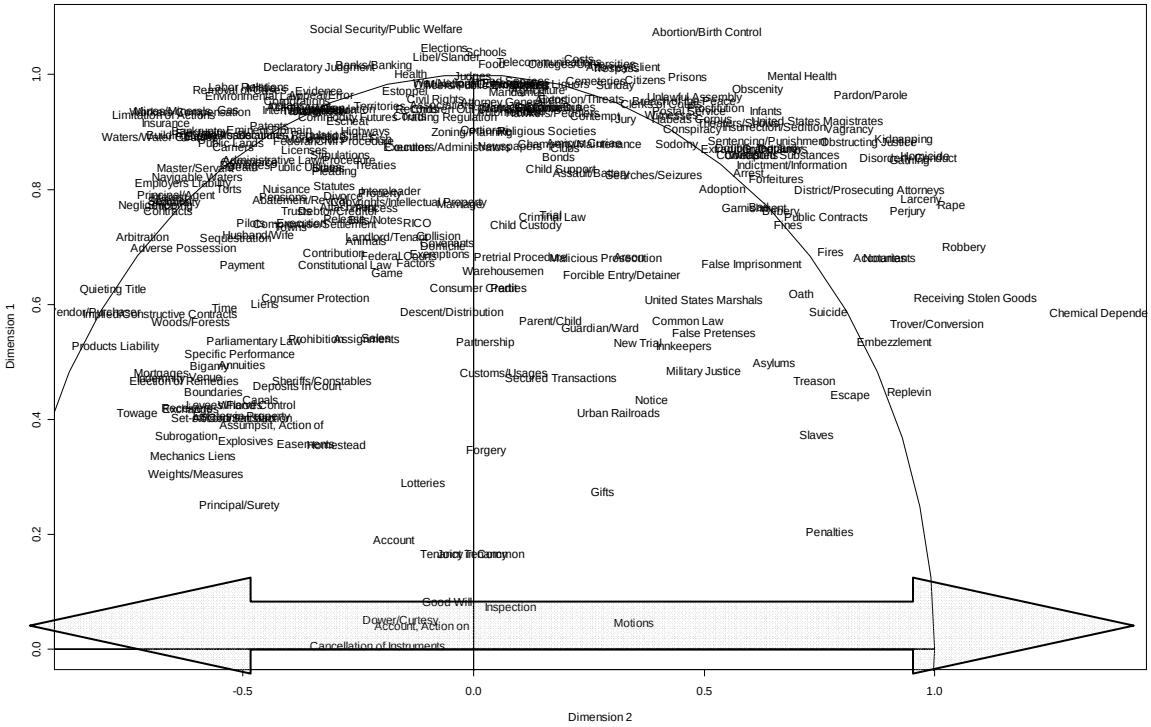
Findings, Discussion and Conclusions

Network Graph Approach

Initial network based attempts to create a domain map of the topic space of Supreme Court cases using the spring force layout algorithm in Pajek proved unsatisfying. The procedural and factual topics, which may co-occur with just about any doctrinal topic, pulled everything to the center of the graph.[2] In order to derive any insight using this approach, the author had to visualize just the doctrinal topics. Furthermore, to obtain readable visualizations, all of the co-occurrences were aggregated up from the West Topic level to the law school subject level (the course offered in law school most likely to teach that particular topic). The graph was then subjected to another double treatment. First, the most tenuous (least numerous) co-occurrences between subjects were discarded. This was a bit subjective and was again informed by the author's familiarity with legal topics. It was necessary because almost every subject co-occurred with Constitutional Law and a few other similarly ubiquitous topics. Second, amongst the remaining subjects, the graph was thresholded at 10 or more case co-occurrences between the subjects. This resulted in network visualization pictured in **Figure 1**.

Apparent from the visualization were several counterintuitive adjacencies that reflect the unique jurisdiction of the United States Supreme Court. Maritime cases invoke federal jurisdiction. Furthermore, to the extent that maritime cases involve contract disputes or workers' compensation claims, these issues are heard by the Federal Courts. Outside of the context of maritime law, contracts and workers' compensation cases are state court issues not typically heard by the Federal Courts. Thus, the resultant base map reflects an inherent bias in the dataset. No expert in the law would intuitively co-locate Maritime Law, Workers' Compensation, and Contracts outside the unique context of cases being heard in the Supreme Court.

[2] One reviewer noted the similarity of the problem encountered by Small and Griffith. In the reviewer's own words, this problem was "the effect of methods papers on document co-citation clustering/mapping (these must be removed before a structure can be found -- see any number of papers by Small on this)." (*See* Small & Griffith, 1974). I wish to thank both unknown reviewers for their comments and feedback.



Business Matters (Property) Criminal Matters (Life & Limb)

Figure 2: PCA (2nd and 1st Dimensions)

PCA Approach

A plot of the amount of the variance contained in each of the singular values reveals that the first twenty-five dimensions account for almost 4/5th's of the variance. On the whole, the dimensionality plots do not reveal easily identifiable continuums. However, the plot of the 1st and 2nd principal components reveal a readily identifiable continuum between criminal matters on one end (Receiving Stolen Goods, Rape, Robbery, Larceny, Homicide, etc.), and business matters on the other (Quieting Title, Constructive Contracts, Mortgages, etc.). This division between matters of life and limb and those of property corresponds with the popular perception of the justice system as being composed largely of two parts—criminal and non-criminal matters. **See Figures 2.** This same continuum may also be seen in a non-PCA layout of the topic relationships of one particular Supreme Court term (2004). Cases as nodes are linked to the topics they contain which are also portrayed as nodes. The spatial layout was generated by hand employing the heuristic charge to minimize edge crossings. **See Figure 3.**

The layout of topics of the first two principal components revealed topic adjacencies that are contrary to traditional categorizations. For instance, the topic Bigamy, which is a crime, appears on the Business Matters end of the previously identified continuum. This at first appears to be an error. However, further research reveals that the topic Bigamy appears only once in the entire dataset. It occurs in the context of a divorce case in which alimony and the division of marital property were hotly contested. In fact, the alleged bigamy (one spouse got a divorce and remarried in a different state and these actions were not recognized by the original state) was the means to the end of acquiring more marital assets in the divorce proceeding. Thus, the appearance of the topic Bigamy at the Business/Property side of the continuum makes sense even though it is contrary to how a law student would encounter the topic. **See Figure 4.**