CS224W: Final Report - Discovering interesting structures and patterns within legal citation networks

David Chin-lung Fong, Felix Long-yin Yu (Group 24) {CLFONG, FELIXROR}@STANFORD.EDU

Abstract

With citation as links, legal cases form a directed graph with a rich set of metadata on each node. In this paper we attempt to uncover interesting structures and patterns within the citation network. For instance, what are the important legal concepts governing the legal opinions, how to identify important legal cases and what are their characteristics. are there any significant clustering patterns within the citation network, etc. We used a wide range of tools to help us achieve our aforementioned goals. The tools employed including PageRank, HITS, tf-idf, Latent Semantic Indexing, and Spectral Graph Partitioning.

1. Overview

Legal cases formed a network linked together by citation. In this project, we employed various graph analysis algorithms to find interesting structures in this network. Tools we used including running PageRank and HITS algorithms to uncover hubs and communities, hence finding cases with high authorities, spectral graph partitioning to look for clusters in the network. We also used machine learning techniques such as Latent Semantic Index and tf-idf to identify important legal concepts.

1.1. Dataset

Our dataset is provided by Andrew Baine from Stanford Law School. It consists of 6.5 million legal opinions from the United States Judiciary from 1850 to the present. Each legal opinion is in the form of an XML document, with markups for various metadata of the opinion. The legal opinions are linked by how they have cited each other, forming a directed graph that could be used for analysis by various algorithms.

1.1.1. Adjacency Matrix

Before running an algorithmic investigation of the legal citation network, we first run a visual inspection of the distribution of non zeros in the adjacency matrix. For the sake of better comparison with the results of spectral graph partitioning, we assume the graph being undirected here. (The graph will be treated normally as a directed graph for PageRank and HITS computation.)

From Figure 1, we see the dataset consists of legal opinions ordered by which state of the court or the level of the court if it is a federal court. Within courts coming from a certain state, there is a high density of citation among each other. This can be confirmed by the solid blocks along the diagonal in the figure. In contrast, courts from different states rarely cite each other. This is confirmed by the fact that off-diagonal blocks in plot exhibit very sparsely distributed non zeros.

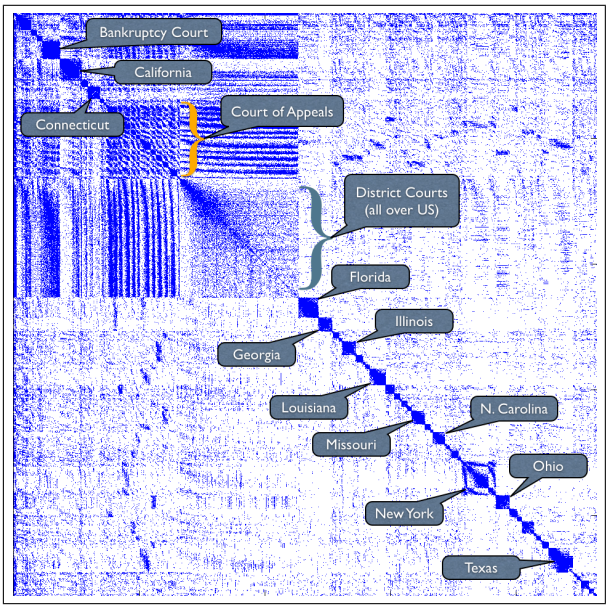

Figure 1. Visualization of non-zeros distributions in the adjacency matrix of the legal citation network

2. Methodology

2.1. Identifying important legal concepts

Legal opinions are of less value to a computer retrieval system when it is in raw text form. Important keywords or concepts can be generated by machine for better retrieval. Keywords that appear more frequently in a particular document than it's overall appearance in the whole collection exhibit a higher significance for that document. These keywords, when extracted from legal opinions, will give a representative summary of that opinion. Furthermore, since the number of possible keywords is huge (most of English words except stop words), a further step for clustering and classification of documents is to run latent semantic indexing [2] on the keyword-document incidence matrix, and extract the compressed concept space of legal opinions.

2.1.1. Extracting words from raw text

The raw data set is a collection of XML documents with markups on variable attributes of each legal opinion. The XML markups are removed and tokenize by splitting on all non-alphanumeric characters. This gives a collections of words that appeared in each legal opinion.

2.1.2. Estimating the importance of each word

First, stopwords are removed using a list of stopwords. Then we compute the tf-idf [5] weight of each word in each document. Other research have shown that this gives a reasonable evalution on the importance of each word.

2.1.3. Latent semantic indexing

The tf-idf weights computed forms a sparse matrix of keyword-document relationship of the collection of legal opinions. We use singular value decomposition to perform latent semantic indexing (LSI) on this matrix, and find the concepts extracted by LSI corresponding to the top singular values.

2.2. Finding cases of high importance

Since legal opinions are linked, the associated citation structure can be exploited to find legal cases of high importance. Similar to how web search engines are built, link analytic algorithms like PageRank [1] and HITS [6] can be used to determine whether a case is of high value relative to other cases. We can then observe if there are any interesting observations and commonalities regarding cases with high PageRank and HITS score.

2.2.1. Appropriate sampling of data

As we can see from the clustering pattern of the raw adjacency matrix in Figure 1, legal cases from different states rarely cite each other. Hence, in order to sample a subset of data that best preserves the underlying link structure of the original dataset, we shortlisted 650000 cases (1/10 of the full dataset) that includes all cases from the states Alaska, Arkansas, California, and Connecticut. The citation structure of these cases should assemble that of the full citation network.

2.2.2. Extract cases with top PageRank and HITS score

PageRank and HITS algorithms associate each legal case with a score that somehow measures how important the case is with respect to rest of the cases. We extracted the top 10 cases for PageRank and HITS score respectively and observed if there are any interesting observations and commonalities regarding cases with high PageRank and HITS score.

2.3. Spectral graph partitioning

To analyze if there are hidden clusters in the citation network that are not revealed by the raw adjacency matrix, we ran spectral graph partitioning on various subgraphs on the dataset. In particular, we have tried to partition on:

- First 10000 nodes
- Random samples of nodes
- Using cases from a single state

To compute the eigenvector corresponding to the second smallest eigenvalue for spectral graph partitioning, we use ARPACK [7] to run Arnoldi iterations. This allow us to scale up the computation to work on graphs with about 200000 nodes. In addition, to ensure scalability, we also used EIGIFP [4] to perform eigenvector computation.

3. Experimental Results

3.1. Identifying important legal concepts

In this experiment, we sampled a subset from the whole data set to perform LSI. A total of 6700 documents are used in this analysis.

The concepts corresponding to the top singular values are shown in Table 1. By inpecting the concept table, we found that Concept 1, which correspond to the largest sigular value, consists mainly of keywords that are common to most of the legal opinions, such as 'case', 'court' and 'defendant'. As we go through the concepts in the order of decreasing sigular value, we see more specific keywords for different types of court cases, such as 'petitions', 'certiorari'. We also found some geographically specific keywords, which shows the regional segmentation of our dataset of legal cases, which further reinforce our observation in distribution of non zeros in the adjacency matrix, where geographical difference across legal cases are the strongest signal to partition the cases.

3.2. Finding cases of high importance

3.2.1. Distribution of PageRank and HITS scores

We performed link analytic algorithms like PageRank and HITS to uncover legal cases of high importance. Figure 2 shows that the distribution of PageRank follow Power Law. Such result makes sense since legal citations should follow the "rich get richer" phenomenon characterized by Power Law Networks in the sense that more cited cases inherit higher tendency to get cited in the fugure. PageRank, a heuristic measure of how often a particular legal case is cited, should therefore follows Power Law.

Similarly, Figure 3 and 4 show that the distribution of Hub and Authority scores also follow Power Law, since their underlying idea is very similar to that of PageRank in the sense that more cited cases are more likely to get cited, and that cases that cite highly cited cases inherit high Hub score.

It is noteworthy to see from Figure 5 that the raw Degree distribution of the legal citation network does not follow Power Law. Performing a linear regression fit to it yields a fitting coefficient of **1.54** that is below 2, the Power Law coefficient that is frequently observed in real world Power Law network. One possible explanation is that PageRank and HITS both take into account the quality of the linkages in the sense that PageRank and HITS tries to capture how much a case is being cited by important cases. But raw degress does not capture such extra dimension.

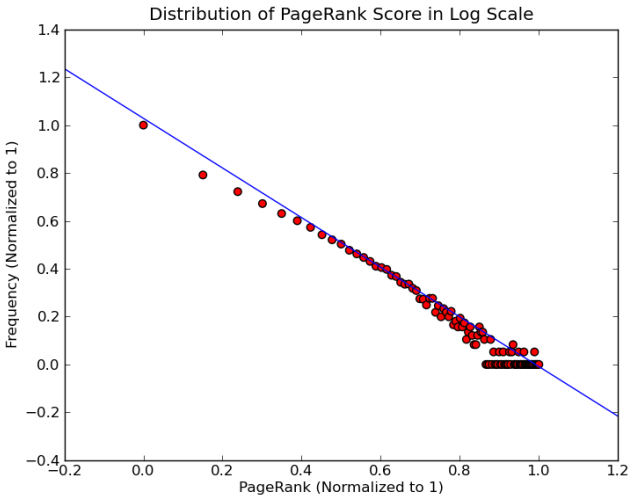

Figure 2. Distribution of PageRank score in logarithmic Scale

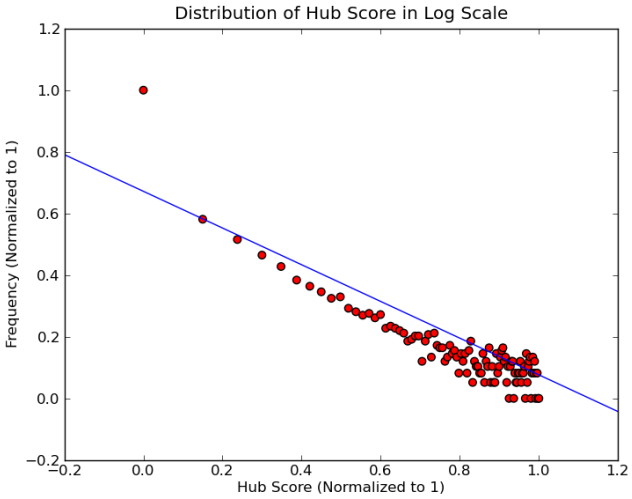

Figure 3. Distribution of Hub score in logarithmic Scale

3.2.2. Cases with top PageRank and HITS scores

We extracted the top 10 high scoring cases for PageRank, Hub, and Authority scores respectively. The top 10 cases for PageRank is listed in Table 2. We can see that the majority of them are old cases that come before 1920. It makes sense since on the one hand these cases are heavily cited by newer cases, hence obtain high in-links. On the other hand, they do inherit many out-links since there are not many cases come before them.

Table 3 and 4 refer to the top 10 cases for Hub and Authority scores respectively. We can see that all of them are from the Supreme Court of California. In fact, we would expect cases with high HITS score are from

	Concept 1	Concept 2	Concept 3	Concept 4	Concept 5
positive weight keywords	case	decided	cir	spurgeon	dept
	court	1st	appx	shaia	super
	defendant	div	writ	majette	pasuper
	may	nycase	donahou	pratts	applications
	order	leave	people	freidinger	petitions
	quot	applications	writs	people	com
	state	york	reported	petitions	nyapp
	supreme	dept	favreau	certification	certiorari
	true	nyapp	uscase	njcase	people
	upon	people	certiorari	jersey	uscase
negative weight keywords	people	certiorari	jersey	com	florida
	nyapp	uscase	njcase	pasuper	fla
	dept	petitions	certification	super	flcase
	york	writs	com	pennsylvania	dist
	certiorari	favreau	super	wda	published
	leave	reported	pasuper	eda	app
	applications	jersey	petitions	superior	mass
	div	njcase	pennsylvania	hoffer	commonwealth
	uscase	writ	wda	mda	curiam
	nycase	certification	eda	pacase	macase

Table 1. Concepts extracted by latent semantic indexing corresponding to the top singularvalues

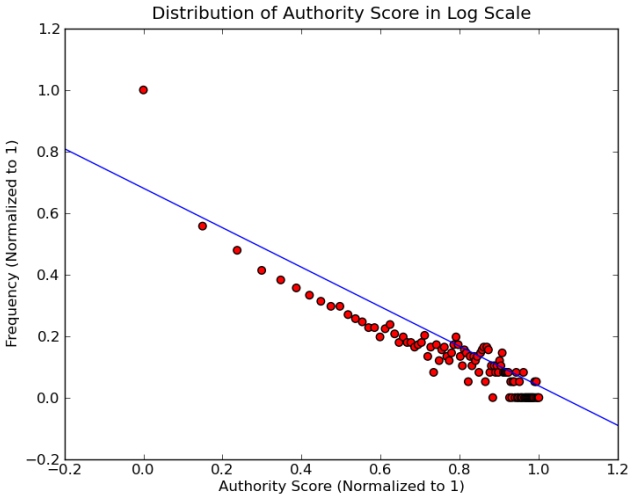

Figure 4. Distribution of Authority score in logarithmic Scale

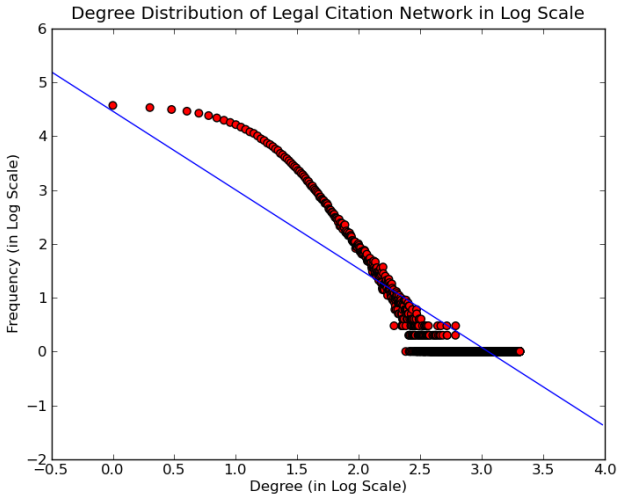

Figure 5. Degree Distribution of the legal citation network in Logarithmic Scale

big States since cases across different States rarely cite each other, so cases from big States form a large cluster and accumulate a relatively large number of citations compared to cases from smaller States. Since California is the biggest State in our sampled dateset it makes sense that the top 10 cases for HITS score are from California.

3.3. Spectral graph partitioning

Several different subgraph of the citation network are extracted for spectral graph partition analysis. Here are the detail results.

3.3.1. First 10000 nodes

From Figure 6, we see that spectral graph partitioning pushed the non zeros closer to the main diagonal. However, the effect is not very significant and no

	Legal Case
1	EVANS v. EVANS, 200 Ala. 329 (1917)
2	HUBBARD v. N. Y., N. H. & H.R. CO., 72 Conn. 24 (1899)
3	ADAMS v. WALSH, 200 Ala. 140 (1917)
4	MIDDLETON v. ARK. EMP. SEC. DIV., 265 Ark. 11 (1979)
5	FRANK v. PICKENS & SON CO., 264 Ark. 307 (1978)
6	LEGNARD v. PLANNING AND ZONING COMMISSION OF BETHEL, 181 Conn. 753 (1980)
7	BAYON v. BECKLEY, 89 Conn. 154 (1915)
8	FAIRFIELD COUNTY NATIONAL BANK v. HAMMER, 89 Conn. 592 (1915)
9	STAPLES v. HENDRICK, 89 Conn. 100 (1915)
10	ALLEN v. JACOB DOLD PACKING CO., 204 Ala. 652 (1920)

Table 2. Top 10 cases by PageRank

	Legal Case
1	PEOPLE v. BRADFORD, 15 Cal.4th 1229 (1997)
2	PEOPLE v. CARPENTER, 15 Cal.4th 312 (1997)
3	RENEE J. v. SUPERIOR COURT, 26 Cal.4th 735 (2001)
4	PEOPLE v. BARNWELL, 41 Cal.4th 1038 (2007)
5	PEOPLE v. SUPER. CT. OF LOS ANGELES CO., 18 Cal.4th 667 (1998)
6	CARMEL VALLEY DISTRICT v. STATE, 25 Cal.4th 287 (2001)
7	TOWNSEL v. SUPERIOR COURT, 20 Cal.4th 1084 (1999)
8	MAYNARD v. BRANDON, 36 Cal.4th 364 (2005)
9	PEOPLE v. WARNER, 39 Cal.4th 548 (2006)
10	PEOPLE v. FRENCH, 43 Cal.4th 36 (2008)

Table 3. Top 10 cases by Hub

	Legal Case
1	PEOPLE v. WATSON, 46 Cal.2d 818 (1956)
2	WOODLAND HILLS RESIDENTS ASSN., INC. v. CITY COUNCIL, 26 Cal.3d 938 (1980)
3	PEOPLE v. HOUSTON, 42 Cal.3d 595 (1986)
4	SOUTHERN CAL. GAS CO. v. PUBLIC UTILITIES COM., 38 Cal.3d 64 (1985)
5	IN RE PEDRO T., 8 Cal.4th 1041 (1994)
6	MARIA P. v. RILES, 43 Cal.3d 1281 (1987)
7	ROSENTHAL v. STATE BAR, 43 Cal.3d 658 (1987)
8	ADOPTION OF ALEXANDER S., 44 Cal.3d 857 (1988)
9	MARDIKIAN v. COMMISSION ON JUDICIAL PERFORMANCE, 40 Cal.3d 473 (1985)
10	PEOPLE v. BALDERAS, 41 Cal.3d 144 (1985)

Table 4. Top 10 cases by Authority

strong clustering effect is found to enable partitioning.

3.3.2. Randomly sampled nodes

With 200000 randomly selected nodes, ARPACK took about 5 hours to converge to a pretty loose tolerance of `1e-5`. From Figure 7, we see the nodes are pushed strongly towards the main diagonal. There some level of clustering effects exhibited, as shown by the white crosses that appeared in the figure.

3.3.3. Florida State

Given the strong clustering by states as shown in the raw adjacency matrix, we hypotensized that if we focus on a single state, we might be able to partition cases based on the legal disciplines represented by different cases. However, when we test this hypothesis using several different states, we found no internal clustering that could be exploited by spectral graph partitioning. One typical example of this is Florida, which is shown

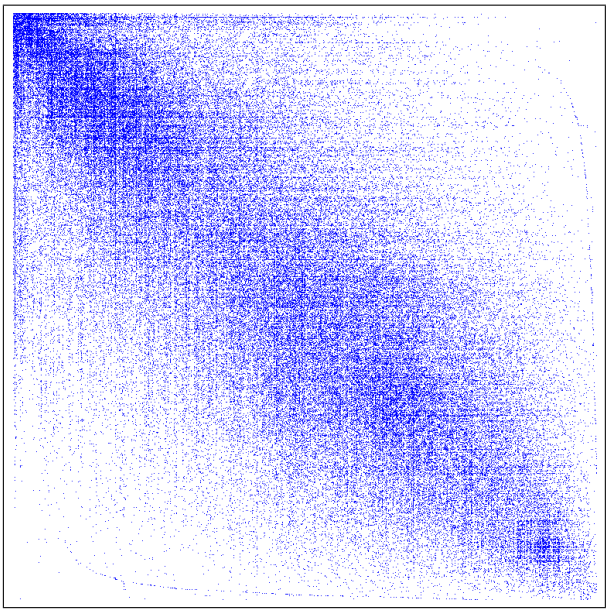

Figure 6. Non zero distribution of first 10000 nodes of the network with spectral graph partition

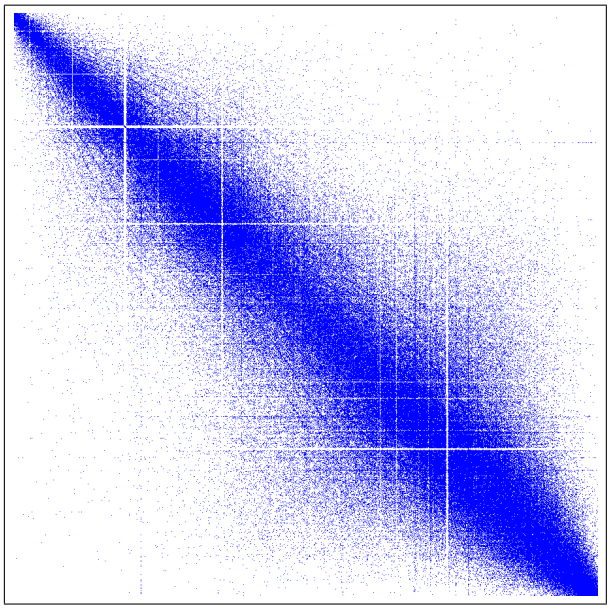

Figure 7. Non zero distribution of 200000 randomly sampled nodes of the network with spectral graph partition

in Figure 8.

4. Future Directions

If possible, we would like to run all the aforementioned algorithms on the full dataset instead of just a subset of the data. Although we already tried various approaches to sample the data such that the sampled data, given their sizes, can best capture the underlying structure of the full dataset, it would be even better if we can run our algorithms on the full dataset and see if our conclusions still apply.

Futhermore, we would like to come up with methods to classify cases as pro-plaintiff or pro-defendant, and see if there are any cascading patterns with regard to court decisions within the citation network. However, doing so will require labeling of existing cases. Upon looking into the case documents we have yet to come up with a reasonable heuristic to successfully label the cases. In addition, it is not easy to assign binary labels on cases since court decisions very often spans a wide spectrum (for example, it is hard to determine if 3 year of imprisonment is favorable for the defendant due to the severity of his or her crime commitment).

5. Conclusion

Given a citation network of 6.5 million legal cases, we adopt various approaches to discover interesting structures and patterns within the network. In particular, we employed Latent Semantic indexing to ascertain important legal concepts, and identified some keywords that are coherent to our context. We also ran link analytic algorithms like PageRank and HITS and discovered the distribution of PageRank and HITS score follow Power Law. In addition, we made some interesting observations regarding the top scoring cases in PageRank and HITS. We observe that the top cases in PageRank are mostly old cases while the top HITS cases are exclusively from big states like California. Last but not least, we peformed Spectral Graph Partitioning on the network but ascertain no significant nternal structuring pattern. Looking forward, we hope to extrapolate our algorithms on the full dataset and also be able to classify cases as pro-plaintiff and pro-defendant given their context in relation to the link structure of the citation network.

References

[1] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Computer Networks and ISDN Systems*, volume 30, pages 107–117, 1998.

[2] S. Deerwester. Improving Information Retrieval with Latent Semantic Indexing. In *Proceedings of the 51st ASIS Annual Meeting (ASIS '88)*, volume 25, Atlanta, Georgia, October 1988. American Society for Information Science.

[3] M. Fiedler. Algebraic connectivity of graphs.

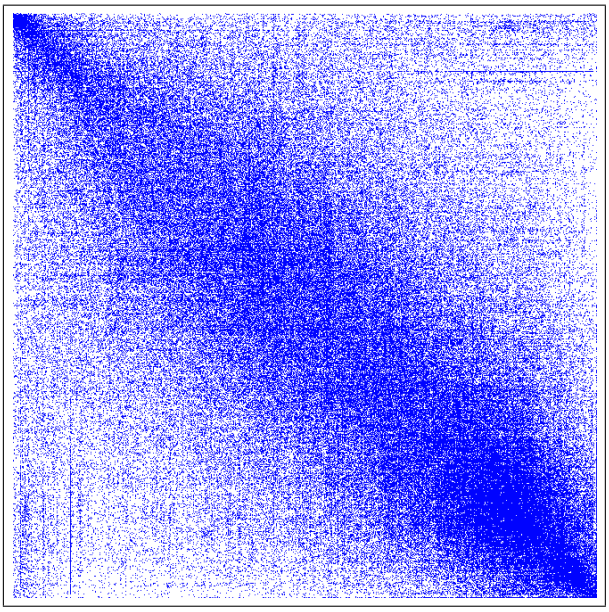

Figure 8. Non zero distribution of florida cases with spectral graph partition

Czechoslovak Mathematical Journal, 23(98):298–305, 1973.

[4] G. H. Golub and Q. Ye. An inverse free preconditioned krylov subspace method for symmetric generalized eigenvalue problems. *SIAM J. Sci. Comput.*, 24:312–334, January 2002.

[5] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[6] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46:668–677, 1999.

[7] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK User's Guide: Solution of Large-Scale Eigenvalue Problems With Implicityly Restorted Arnoldi Methods (Software, Environments, Tools)*. Soc for Industrial & Applied Math.