USING LATENT DIRICHLET ALLOCATION TO EXPLORE THE DIMENSIONALITY OF THE U.S. PRACTICE OF LAW

Patrick H. Gaughan,[1] En Cheng,[2] Taylor C. Burgess,[2]
and Aine C. Bolton[2]

[1]School of Law, University of Akron, Akron, Ohio, USA
[2]Department of Computer Science, University of Akron, Akron, Ohio, USA

ABSTRACT

Over the centuries, the U.S. practice of law has evolved into a complex and amorphous profession. To facilitate improved analysis and understanding, this exploratory study seeks to partition law practice areas into meaningful subgroups. The study applies Latent Dirichlet Allocation ("LDA") as a soft clustering method to 437,210 individual U.S. lawyer profiles in private practice in 2000. The profiles came from a nationally recognized directory. The resulting subgroupings contain terms consistent with the hypothesized relationships. The results also suggest the possibility of systematically binning individual practice areas into discrete practice area distributions. As such, this study makes contributions to the existing literature in at least three areas: 1) it provides support for the existence of the hypothesized law practice relationships; 2) it provides an empirical basis for developing an improved measurement of the U.S. practice of law; and 3) this study also suggests additional research to advance the field.

KEYWORDS

Law Practice, Practice Area, Practice Dimensionality, Practice Partitioning, Practice Grouping

1. INTRODUCTION

For centuries, the U.S. practice of law has been conceived of as being homogenous. The taken-for-granted assumption has been that U.S. lawyers engage in the singular "practice" of law – not "practices" of law [1][2]. Consequently, many studies of the U.S. "practice of law" simply segment U.S. lawyers based upon demographics. See, e.g. [3]. More nuanced studies rely upon dissociated, often self-defined, practice areas. See, e.g. [4]. Across all of these studies, there is little effort to investigate differences within the practice of law.

At one time, the ambiguities did not matter. The U.S. practice of law was less complex than it is today. Prior to the 1870's, U.S. lawyers received an apprentice-like education and tended to provide generic legal services [5]. But as the practice of law became more complex, the U.S. legal profession continued to embrace the singular conception of "the Law" and "the practice" of law [2].

Compounding the problem, U.S. courts have habitually provided vague definitions of what exactly constitutes "the practice of law" [6]. For instance, one U.S. Supreme Court decision defined lawyers as persons "acting professionally in legal formalities, negotiations, or

David C. Wyld et al. (Eds): IOTBS, NLTM, AIMLA, DBDM -2024
pp. 43-57, 2024. CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.140204

proceedings by the warrant or authority of their clients…" [7]. Similarly, one state supreme court recently pronounced that "a person engages in the practice of law when he counsels or assists another in matters that require the use of legal discretion and profound legal knowledge" [8].

Making empirical study even more complicated is the fact that the provision of legal services involves more than just supply-side considerations [9][10]. Today, extensive heterogeneity exists as to the sophistication, resources, experience, expectations and needs of law clients. For instance, there are widely recognized differences between corporate and individual clients – even within nominally similar law practice areas [11]. Additionally, legal services tend to possess an especially opaque quality [12][13]. As such, some clients have difficulty evaluating the composition of legal services prior to their delivery [12]. Therefore, subtle differences may exist between semantically similar practice areas that reveal less about word meanings than about different service qualities (as intended for different potential client groups).

Accordingly, the challenge is to partition law practice areas "into meaningful subgroups, when the number of subgroups and other information about their composition… is unknown" [14]. With the proper method, each subgrouping of practice areas then constitutes a different element within a single dimension [15]. This can then be used for binning the practice areas (as categorial variables) in discrete buckets. [16][17]. This technique has previously been used to categorize other types of information regarding the U.S. practice of law. [18].

In order to do this, the present study used a dataset extracted from a national directory of U.S. lawyers in private practice in 2000. After data cleaning and setting a minimum frequency for each practice area, the study included profiles from 437,210 individual lawyers. These profiles contained a total of 1,058,788 practice area entries. The study then used Latent Dirichlet Allocation as a soft clustering method [19][20].

The findings of this study suggest that the hypothesized optimal dimensionality exists. These results can now be used as a means of grouping practice areas into "elements of the same type" [15]. In turn, this grouping will enable the informed investigation within the U.S. practice of law.

2. HYPOTHESIS DEVELOPMENT

Although often tempered by appeals to professionalism [21]. and avoiding prejudice to the administration of justice [22], most U.S. lawyers develop their law practices at least partially based upon economic considerations. Lawyers make conscious decisions regarding the specific legal services to provide and where to provide them [23]. However, most lawyers face inherent challenges in aligning their service/location decisions with the external market for their legal services.

As with other services, legal services tend to be intangible, nonstandard, and inseparable from production and consumption [24]. However, legal services are also "credence goods" with an especially opaque quality [25]. Within any given practice area, many clients have varying degrees of sophistication, resources, experience, expectations and needs. Many clients may also have difficulty comparing the services offered by competing lawyers [26] This means that many potential law clients have little choice but to look to a lawyer's past history and reputation when considering a potential relationship [27]. More generally, potential clients may also look to the reputation of the law firm where the lawyer works [28]. However, in both instances, reputation signals to potential clients how the legal services may compare to competing firms [29].

Given the credence characteristics of legal services, client-side perceptions have an impact on how U.S. lawyers develop their law practices [10]. As a lawyer gets experience in a given

practice area, the lawyer will begin developing a corresponding reputation. In the process of building strong client relationships, the lawyer develops competitive advantages over other lawyers providing similar legal services [30]. However, the development of a reputation in one client-service segment can be both a blessing and a curse.

A positive reputation will make it easier for the lawyer to get new clients of the same type as that of their existing clients [30]. However, an emerging positive reputation in a specific area may tend to type-cast the lawyer. This may have the effect of excluding the lawyer from other areas perceived by potential clients as being too remote. In effect, the lawyer will tend to obtain new clients in an area narrowly defined by the lawyer's past work or otherwise related to the needs of existing clients [31]. Even though reputational information is not fool-proof [32], an emergent career path dependency will develop as a lawyer's reputation grows and more potential clients become aware of the lawyer's existence [33].

Based upon the above-listed theoretical mechanisms, the likely co-occurrence of two or more practice areas within any individual lawyer profile will not be randomly determined. Some term co-occurrences will reflect the supply-side (lawyer-based) preferences for semantically similar practice areas. At the same time, other term co-occurrences will reflect dissimilar practice areas that are commonly offered together based upon the demand-side (client-focused) needs of the legal services market.

This gives rise to the following hypothesis:
H_0: The distribution of U.S. practice areas (across individual lawyer profiles) will be randomly dispersed without any indication of any meaningful grouping.
H_A: The distribution of U.S. practice areas (across individual lawyer profiles) will not be randomly dispersed but will contain apparent indications of some meaningful grouping.

3. METHODOLOGY

One of the most popular, unsupervised, probabilistic topic modelling techniques is Latent Dirichlet Allocation ("LDA"). See [34][35][36]. An initial concern with using LDA in this study was that all of the 437,210 documents were exceptionally short. On average, each document consisted of less than three terms. This raised the prospect that the document-level word corrections might be too low [37]. However, unlike other LDA studies that analysed tweets, the documents used in the present study contained only the most distilled, critical, information of interest. Additionally, as explained more fully below, these observations consisted of an extremely limited variety of terms. After setting the minimum frequency of 0.001 of all initial observations, there were only 100 unique practice area terms. In effect, LDA was used for soft clustering of 100 terms within a document probability model of 437,210 documents and 1,058,788 total terms. As a result, it was believed that LDA was appropriate for the intended task and any non-zero probabilities, even if small, were likely to be significant.

An additional concern regarded the limitations of potential alternative methods like the Biterm Topic Model. Unlike LDA, the Biterm Topic Model "explicitly models the word co-occurrence patterns in the whole corpus to solve the problem of sparce word co-occurrence at the document level" [35]. While the Biterm Topic Model addresses the problem of sparce data, it does so by analysing co-occurrences across the entire corpus rather than within documents. Even setting the Bitermplus screening window to the average document length of just three (3) would still enable the Bitermplus to calculate term co-occurrences across documents. Given these concerns, the authors deemed the LDA method to be the most appropriate for the present study.

In preparing the dataset for processing, the basic methodology was patterned after the process overview contained in Asmussen, C.B. and Møller, C [38]. This included: pre-processing the corpus and the dictionary; setting the parameters for LDA; and cross-validation of the results. [38]

3.1. Pre-Processing: Load the Corpus

The initial data used to constitute the corpus for analysis came from the Martindale-Hubbell Law Directory on CD-ROM, dated Summer, 2000 (the "Martindale-Hubbell"). For analytic purposes, the only entries that were searched were from the main category of "US Lawyers and Firms (Disc 1)." This therefore did not include lawyers practicing outside the U.S., corporate attorneys, government attorneys or faculty.

After selecting the "Advanced Search" option in the CD-Rom containing the 2000 Martindale-Hubbell Law Directory, the radio button for "Lawyers Only" was selected. Clicking the "Search" button revealed a total of 739,458 initial lawyer profiles. By comparison, the ABA estimates that the total number of ALL U.S. lawyers in 2000 (including corporate attorneys, government attorneys and faculty) was 1,022,462 [39]. The sample frame was exceptionally comprehensive.

After extracting the name and practice area profiles from the Martindale-Hubbell CD-Rom, the 739,458 profiles were filtered to remove 252,495 entries that contained blank "Practice Area" fields. This left 486,963 records. Lastly, records suspected of containing duplicate attorney names were removed. This resulted in 454,660 unique records, all with non-blank "practice area" fields.

3.2. Pre-Processing Corpus: Clean Documents and Validate Cleaning

Preliminary visual examination of the 454,660 records revealed that the vast majority of "Practice Areas" consisted only of a list of nouns or compound nouns separated by commas. However, upon closer examination it was discovered that the formatting appeared to be based solely upon convention. No restrictions had been imposed on how the data was actually input. Some profiles contained non-standard terms, special characters, punctuation errors, inconsistent spacing, dashes, standard and nonstandard abbreviations, and typographical errors. A few records were written as a narrative rather than as a list.

To the extent possible, all formatting issues were addressed as part of the initial clean-up of the Corpus. Cleaning proceeded by removing all short connecting words (like "a," and "the"), and qualifying phrases (like "but not limited to").

Additionally, given the lack of data entry controls, inconsistencies were identified as to the use of singular and plural forms of some practice areas. For this reason, all key words (both in the dictionary and the corpus) were adjusted so that only the singular form existed. These changes were made to all nouns and to all components of compound nouns. For instance, some documents referenced "products liability" while other documents referenced "product liability." The string "products" was changed to "product" in both the corpus and dictionary.

Beyond the singular/plural inconsistency, some documents were identified where "Law" was either included or omitted. For instance, some documents would state "bankruptcy," while other documents would state "bankruptcy law." Since the commas were going to be used as delimiters to search individual practice areas, the word "law" was removed from the entire corpus to facilitate the capture of identical strings. Similarly, descriptive terms like "plaintiff" or "defendant" were removed to better capture identical practice areas.

However, given concerns about biasing the results by only partially correcting a subgroup of entries, the decision was made to abstain from any other changes to the corpus documents. Different practice areas that appeared to be synonyms, hypernyms, and hyponyms were left unadjusted. No effort was made to reconcile different Practice Area terms with similar semantic meanings. For instance, some documents referenced "intellectual property" while other documents referenced "patents, trademarks, copyrights." Given the difficulties of objectively determining where to start and stop, the issue of unaddressed semantic similarity was left for future research.

Additionally, the corpus was adjusted to align with pattern adjustments made in the dictionary. (See the next section regarding the pre-processing of the dictionary for additional information.) Except for these corrections discussed above, the corpus was not otherwise corrected. No effort was made to modify any typographical errors, or non-standard Practice Area terms. Again, the concern was the inability to determine when corrections should begin or end – and the concern that such corrections might bias the results of the LDA.

As explained in the following section, once all of the documents in the corpus were cleaned, each Practice Area appearing in each document were bigrammed so as to facilitate matching with the terms of art contained within the corresponding dictionary.

3.3. Pre-Processing Dictionary: Cleaning and Validating

Having completed the pre-possessing of the text corpus, the dictionary was assembled by combining the lists of practice areas contained in the appendices to both the 1998 and 2000 Martindale Hubbell Law Directory CDs.

For some unknown reason, the Practice Area list to the 2000 Martindale Hubbell contained only 20 terms. This was in sharp contrast to the 1998 Martindale Hubbell which contained 1964 terms. Given the stark difference in the numbers, and the apparent failure of the 2000 Martindale to control the data field, a decision was made to combine the 2000 and 1998 Martindale practice areas. The concern was that lawyers in the 2000 Martindale might have simply copied profiles from previous years. This could lead to the unnecessary omission of sizable numbers of practice area entries.

First, all entries from the 2000 Martindale Hubbell Practice Area List were retained. Next, all entries in the 1998 Martindale Hubbell Practice Area List that clearly related to any of the 2000 Martindale Hubbell Practice Area List were changed to the 2000 value. Any remaining 1998 Martindale Hubbell Practice Area List were retained in their original value.

As mentioned in the previous section, visual inspection of the combined practice area list revealed a number of areas using either singular or plural terms. As with the corpus, all of the entries in the dictionary were standardized on the singular form (as well as across the corpus). These changes were made to both nouns and components of compound nouns.
Once this was complete, all multi-word terms (all compound nouns) in the dictionary were bigrammed. The logic was that Martindale Hubbell considered each practice area to be a distinct term-of-art.

3.4. Pre-Processing: Set parameters for LDA

The parameter settings for the LDA model were set by applying the information from Rehurek R, 2009 [40]. Similarly, given the separate use of BTM, the parameter settings for the BTM model were set by applying the information from Terpilovskii, M. (2023) [41]. Given the exploratory

nature of the inquiry, the number of topics for both models was set to sequentially increase from 1 to 30. The maximum number of topics was intended to extend beyond any reasonably conceivable expectation regarding the range for potential solutions.

Given the large number of documents in the text corpus, the chunksize was set at 20,000 documents for each training chunk. Passes was set to 1000 for training. Alpha was set to "auto" since there was no a priori belief in the document-topic distribution. "per_word_topics" was set to "true" so that the model would compute a list of topics. "update_every" was set to 1 so that the model would be updated for each document iterated through.

In order to determine the minimum frequency to set for modelling, the frequency of each practice area calculated for all "practice area" terms appearing in both the corpus and list of defined terms. This chart revealed a significant number of very low frequency occurrences.

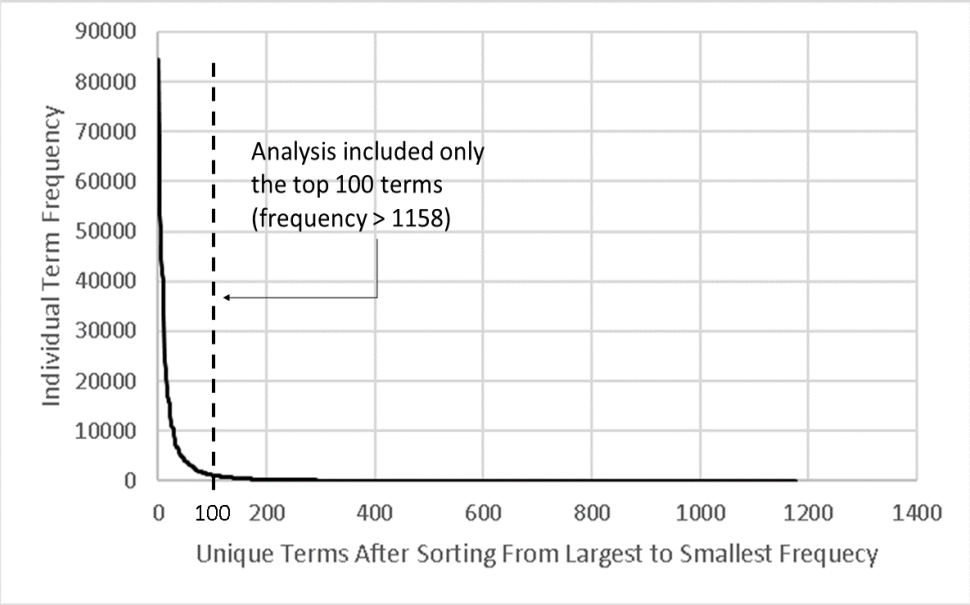

Figure 1: Frequency of all 1176 Unique Practice Area Terms.

For this reason, the decision was made to only include practice area terms with a frequency greater than 0.001 of the total number of original observations, prior to cleaning (1,158,754). This resulted in setting the minimum frequency of terms at 1,158. Applying this minimum frequency resulted in only 100 of the 1176 original terms being included in the analysis. However, as suggested by the chart below, the 100 terms possessing the minimum required frequency accounted for 1,058,788 occurrences out the 1,158,754 original total. This meant that the restricted model was still using 1,058,788 / 1,158,754 = 91.37% of the entire available data. Notably, about 14,000 individual entries did not contain any of the defined practice area terms. Excluding these resulted in 437,210 observations being subject to the LDA analysis.

Looking at the list of the top 100 practice areas, it was immediately clear that there are uncontrolled semantic relationships regarding some of the practice areas. For instance, looking at the top 40 practice areas, below, "Personal injury" is likely a hyponym of "litigation." Similarly, "Estate planning" is a hypernym of "probate."

Table 1: Top 40 Practice Areas by Frequency

Rank	Practice Area Term	Freq.	Rank	Practice Area Term	Freq.
1	Real_Estate	84376	21	Construction	15503
2	Personal_Injury	67423	22	Taxation	13187
3	Commercial	53109	23	Intellectual_Property	12983
4	Corporate	50829	24	Will	11808
5	Business	47339	25	Labor_and_Employment	11036
6	Civil_Practice	44499	26	Trust_and_Estate	10841
7	Criminal	42591	27	Municipal	10670
8	Litigation	41483	28	Administrative	10469
9	Probate	40334	29	Banking	10054
10	Family	37660	30	Trademark	9242
11	Estate_Planning	32994	31	Appellate_Practice,	8404
12	Product_Liability	24363	32	Health_Care	7351
13	Bankruptcy	24004	33	Patent	6896
14	Insurance	22350	34	Merger_Acquisition*	6825
15	Worker_Compensation	20511	35	Contract	6818
16	Insurance_Litigation	18853	36	Professional_Liability	6773
17	General_Practice	17667	37	International	6620
18	Medical_Liability	16828	38	Negligence	6243
19	Environmental	16392	39	Civil_Right	5612
20	Security	15841	40	Collection	5494

* - Indicates that the full practice area name has been shorten due to space limitations.

At the same time, it was felt that the apparent semantic similarities between practice area terms might reflect different socio-semantic network relationships [42]. For instance, “Wills” and “Estate Planning” might be considered to be synonyms. However, the different semantic functions might reflect different types of client groups. For this reason, the decision was made to refrain from any changes to the coding of practice areas based upon any apparent semantic relationship. The decision was made to defer consideration of these options for future research.

3.5. Pre-Processing: Cross-Validation

Due to the exploratory nature of the analysis, the number of potential topic solutions was set to show the cohesion scores of 1 to 30 numbers of topics. Topic solutions in excess of 30 would likely prove to be of little practical value. However, the prospect that there would be maximum coherence scores in models with large numbers of topics was still relevant in determining whether or not there was any particular dimensionality in the data. By placing the maximum number exceptionally high, the research would be better able to provide persuasive support regarding the dimensionality hypothesis.

4. RESULTS AND DISCUSSION

The processing of results followed a somewhat straightforward progression. First, the coherence score was plotted for each topic solution ranging from 1 to 30. Second, based upon the chart of the topic solution with the maximum coherence score, the solution with the best inferred fit was selected. This solution was then used to allocate the practice areas in the corpus into different groups. The groups were then examined for face validity.

4.1. Topic Modelling Overview

Listed below is the chart of the coherence scores for the separate topic solutions ranging from 1 to 30. Even though the maximum coherence score value was slightly over .383, the spike was clear regarding the eight (8) topic solution. Additionally, the reader is reminded that a thesaurus was not used to map apparent synonyms, hypernyms and hyponyms across lawyer profiles. These were left unchanged to reflect both supply-side (lawyer-based) preferences and demand-side (client-focused) needs. This was expected to contribute to a certain amount of ambiguity that would depressing the observed coherence scores.

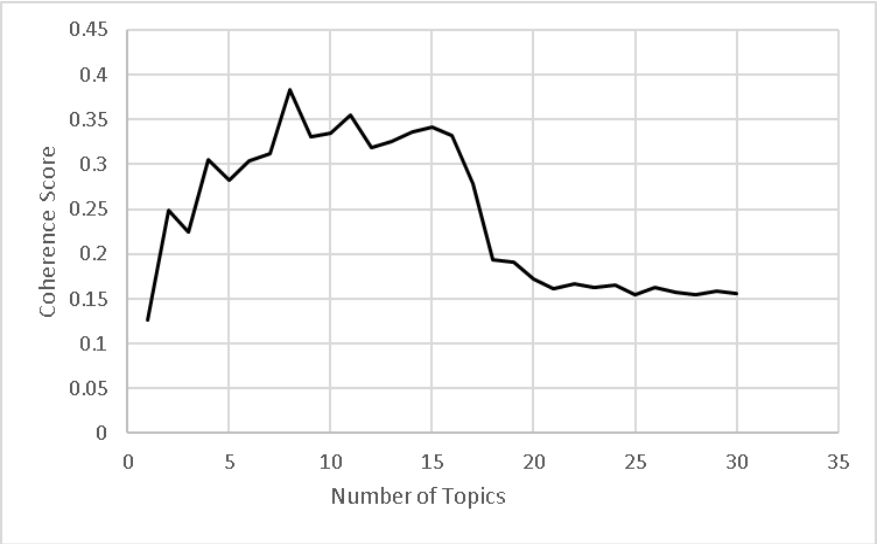

Figure 2. LDA Coherence Scores for Topics 1-30

Based upon the results of the coherence scores listed above, the eight (8) topic solution was selected for closer examination for term grouping.

4.2. Evaluation of Assignments

Once the determination was made to select the eight (8) topic solution, the next step was to quickly look at the terms within each of the groups. Based upon this review, representative names were generated (by the current authors) for each of the topic groups.

Table 2. Practice Group Names for Topics 1 – 8.

Topic 1: Estate Management	Topic 2: Personal Injury
Topic 3: Bankruptcy	Topic 4: Corporate Matters
Topic 5: IP & Criminal	Topic 6: Family Law
Topic 7: Environmental & Government Law	Topic 8: Employment & Mediation/Arbitration

Next, looking at the Interoptic Distance Map, the authors noted the Top-30 most salient terms as well as the total term frequencies within the different topics. This was represented by the size of the respective circles for each topic. The results indicated that Topics 1 and 2 contained the highest frequencies followed by Topics 3 and 4. The interoptic map, using the TSNE setting, is on the following page.

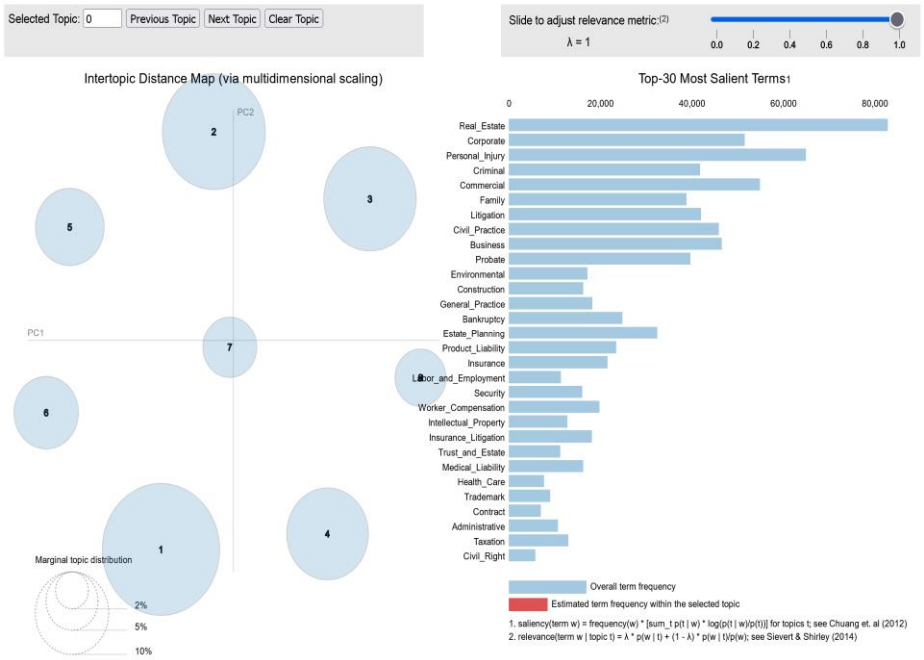

Figure 3. Interoptic Distance Map

Having looked at the broader relationships between the groups, the authors next looked more closely at each of the individual topic groupings. For convenience, the top ten terms (by frequency) within each group are listed below. Additionally, it should be remembered that the groupings are based upon modelling of individual lawyer profiles. As mentioned in Section 2 of this paper, it is hypothesized that the LDA groups reflect both the semantic relationship of lawyer preferences (supply-side) combined with language choices that are responsive to the word choices of unobserved groups of law clients (demand-side) [42]. As such, terms with similar semantic meaning might still be a reflection of different client groups receiving differentiated legal services. Each of the topics is discussed immediately below.

Table 2. The Top Salient Terms for Topics 1 and 2.

Topic 1: Estate Management	**Topic 2: Personal Injury**
Real_Estate	Personal_Injury
Business	Product_Liability
Probate	Insurance
Estate_Planning	Worker_Compensation
Taxation	Insurance_Litigation
Will	Medical_Liability
Municipal	Professional_Liability
Banking	International
Land_Use	Negligence
Commercial_Real_Estate	Tort

For instance, looking at Table 2, Topic 1, “Real_Estate”, “Business”, and “Probate” can be understood as indicating the distinct practice areas that the attorneys in this group tended offer to an unidentified group of clients. Additionally, there are semantically similar terms like “Probate,” “Estate_Planning,” and “Will.” There also are semantically similar terms like “Real_Estate,”

"Commerical_Real_Estate," and "Land_Use." The terms "Municipal" likely relates to "Land_Use." The only questionable term from a casual reading is inclusion of "Banking."

Looking at Topic 2, almost all of the terms directly relate to something connected with "Personal Injury." Even "Insurance" and "Worker_Compensation" are related to either personal injuries or damages of some sort (including personal injuries). The only questionable term from a casual reading is the inclusion of "International."

Table 3. The Top Salient Terms for Topics 3 and 4.

Topic 3: Bankruptcy	**Topic 4: Corporate Matters**
Commercial	Corporate
Civil_Practice	Litigation
Bankruptcy	Security
Administrative	Merger_Acquisition*
Appellate_Practice	Finance
Collection	Employee_Benefit
Creditor_Right	Partnership
Government	ERISA
Foreclosure	Public Finance
Constitutional	

Looking at Table 3, Topic 3, most of the terms seem to revolve around external business relationships and bankruptcy practices. There is a semantic grouping of terms like "Bankruptcy," "Collection," "Foreclosure," and "Creditor_Right". Additionally, the grouping includes semantically related terms like "Administrative" and "Government." There also are more general terms like "Commercial," and "Civil_Practice". However, there is less of a logical connection to the term "Constitutional." This will require additional research.

Looking at Topic 4, Corporate Matters has all of the topics closely linked to the internal operation and/or fund raising activities of business organizations (rather than the external aspects covered in Topic 2). "Security" would include "Securities" [the pre-processing of the data included making all plural forms into the same singular form]. As such, there would be a natural connection with "Finance" and some overlap with "Public_Finance," "ERISA," and "Employee_Benefit." Although "Litigation" is a hypernym for personal injury litigation (see Topic 2), it is also understandable that lawyers addressing "Corporate Matters" would also offer "Litigation."

Table 4. The Top Salient Terms for Topics 5 and 6.

Topic 5: IP & Criminal	**Topic 6: Family Law**
Criminal	Family
Intellectual_Property	General_Practice
Trademark	Trust_and_Estate
Patent	Divorce
Copyright	Child
Matrimonial	Juvenile
Entertainment_and_the_Art	Adoption
Computer_and_Software	
Energy	
Telecommunication	

Looking at Table 4, Topic 5 is the only grouping of terms with an apparent mis-match. The grouping of the term "Criminal" appears to be a mismatch. It does not logically relate to the various other terms related to Intellectual Property such as "Intellectual_Property," "Trademark," "Patent," and "Copyright." Even "Entertainment_and_the_Art" and "Computer_and_Software" have a significant relationship to Intellectual Property. The relationship to terms "Matrimonial," "Energy," and "Telecommunication," are a bit less obvious. These other areas will likely require external investigation to determine the extent to which they below in Topic 5.

Looking at Topic 6, there seems to be a very strong practical relatedness of the different practice areas. Even the inclusion of "Trust_and_Estate" would have a logical relationship to Family Law – even if it also has a strong semantic relationship to Topic 1.

Table 5. The Top Salient Terms for Topics 7 and 8.

Topic 7: Environmental & Government Law	**Topic 8 Employment Law & Mediation/Arbitration**
Environmental	Labor_and_Employment
Construction	Contract
Health_Care	Civil_Right
Anti-Trust_and_Trade	Mediation
Government_Contract	Employment_Litigation
Hospital	Arbitration
White_Collar_Crime	Employment_Discrimination
Natural_Resource	Consumer
Transportation	Education
Class_Action	ADR*

* - Indicates that the full practice area name has been shorten due to space limitations.

Looking at Table 5, Topic 7, there appears to be nice grouping of overlapping areas involving Environmental law and additional related areas with extensive involvement of the government. "Health_Care," and "Hospital" are likely to be significant components of government-intensive practice areas including "Government_Contract," "Anti-Trust_and_Trade," " Transportation," and "Environmental." It is possible that "Construction" was included in the group due to common occurrences with Environmental issues. This leaves only "White_Collar_Crime" and "Class_Action" as somewhat curious members of Topic 7.

Lastly, Topic 8 appears to have closely-related practice areas linked to Employment Law and Mediation/Arbitration (all as a single group) that commonly includes the hypernym "Contract," plus "Labor_and_Employment," and "Employment_Discrimination." Not surprisingly, a closely related area would be "Employment_Litigation." At the same time, Employment Law commonly includes work in "Mediation," "Arbitration," and "ADR." Less obvious is the inclusion of "Education" which may be to collective bargaining, arbitration and labour laws. To a lesser extent, it is understandable that some lawyers practicing in the area of Employment Discrimination would also practice in "Civil Rights."

Overall, the results appear to present subgrouping of terms that are meaningful and consistent with the hypothesized relationships. Most of the term groupings are defensible. However, there are still a few terms presenting minor questions.

4.3. Validity of Allocations

The exploratory nature of this study limits extensive validation. However, the results were subject to face validity evaluation by a licensed lawyer (a co-author). Based upon that evaluation, the groupings appear to be meaningful. The handful of questionable groupings are identified in the prior section.

Beyond the face validity, it should be noted that observed coherence score could be higher than 0.383. The dataset had unaccounted synonyms, hypernyms and hyponyms that would certainly add noise to any LDA analysis. However, no effort was made to control for these issues in order to preserve the potential both the supply-side (lawyer) and demand-side (client) grouping information.

4.4. Limitations and Suggestions for Future Research

Like any other research, this study has some limitations:

1. The short document sizes subject to the analysis (as represented in each individual lawyer profile) could raise some concerns. For LDA, shorter document sizes are not as preferable as having longer documents. However, utilization of a method better suited for short documents (like the Biterm Topic Model) would have deviated from the intended purpose of the study. The current analysis was intentionally focused on practice areas contained within individual attorney profiles – not across the profiles. This is an inherent limitation of the available dataset.
2. The subjective nature of the selected minimum frequency could raise some concerns. Given the sparse data, changes in the minimum frequency for the LDA method could have an impact on the assignment of individual practice areas. It was for this reason that the minimum frequency was determined ex ante to be especially low. Only excluding practice areas with a frequency of less than $1/10^{th}$ of one percent (of all observations) was extremely inclusive. This resulted in the analysis utilizing over 91.7% of the entire available data.
3. The impact of unaccounted synonyms, hypernyms and hyponyms could raise some concerns since they likely reduced the coherence of the LDA output.
4. Possible aspirational bias of in using self-reported practice areas could also raise some concerns. Since all the profiles were self-reported, there is no way to confirm the accuracy of the information that was included in the dataset. Some lawyers may have listed the areas that they wished they practiced in rather than their actual practice areas.

5. CONCLUSION

To the knowledge of the authors, this study is one of the few empirical efforts to attempt the partitioning of the entire U.S. practice of law into meaningful subgroups. The LDA analysis in this study resulted in meaningful subgroupings consistent with the hypothesized relationships. This also suggests the possibility of binning of practice areas into discrete distributions.

ACKNOWLEDGEMENTS

The authors would like to thank: Dr. Rajshekhar (Raj) G. Javalgi for his input regarding the theoretical exploration of law practice dimensionality; Mr. Vasyl Andriiuk for his help troubleshooting the code used in the study; and both the University of Akron (Akron, Ohio, USA) and the Youngstown Business Incubator (Youngstown, Ohio, USA) for their support while

conducting this research. Earlier, preliminary results of this study were presented to the Section on the Empirical Study of Legal Education and the Legal Profession WIP "Incubator" Program on Friday, January 6, 2023, as part of the 2023 AALS (Association of American Law Schools) Annual Meeting in San Diego, CA.

REFERENCES

[1] Zemans, F. K. and Rosenblum, V.G. (1980). Preparation for the Practice of law – the Views of the Practicing Bar. American Bar Foundation Research Journal. 1980(1): 1-30.

[2] Ruan, N. (2014). Student Equire?: The Practice of Law in the Collaborative Classroom. Clinical L. Rev. 20(2): 429-466, 450.

[3] American Bar Association (2022). ABA Profile of the Legal Profession. Available at: https://www.americanbar.org/content/dam/aba/administrative/news/2022/07/profile-report-2022.pdf.

[4] Langford, C.M. (2005). Depression, Substance Abuse, and Intellectual Property Lawyers. Kansas L.Rev.. 53:875-984.

[5] Hurst, J. W. (1967). Lawyers in American Society 1750-1966. Marquette Law Review. 50(4):594-606.

[6] Turfler, Soha F. (2004). Note: A Model Definition of the Practice of Law: If Not Now, When? An Alternative Approach to Defining the Practice of Law. Wash. & Lee L. Rev. 61(4):1903-1959.

[7] National Sav. Bank v. Ward, 100 U.S. 195, 199 (1879).

[8] Dressel v. Ameribank, 468 Mich. 557, 664 N.W.2d 151 (2003).

[9] Gobez, A.R. (2021). Demand Side Justice. Georgetown Journal on Poverty Law and Policy. 28:3: 411-436.

[10] Johnson, V. (1991). On Shared Human Capital, Promotion Tournaments, and Exponential Law Firm Growth (book review), Texas Law Review, 70(2): 537-563.

[11] Campbell, R. (2012). Rethinking Regulation and Innovation in the U.S. Legal Services Market, 9 N.Y.U. J.L. & Bus. 9(1): 1-70.

[12] Darby, M.R. and Karni, E. (1973). Free Competition and the Optimal Amount of Fraud. J. of L. and Econ. 16(1): 67-88.

[13] Ladinsky, J. (1976). The traffic in legal services: lawyer-seeking behavior and the channeling of clients. Law & Society Review, 11(2), 207-224.

[14] Fraley, C. and Raftery, A.E.(1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. The Computer Journal. 41(8): 578-588.

[15] Warfield, J.N. and Christakis, A.N. (1987). Dimensionality. Systems Research 4(2): 69-152, 130.

[16] Verster, T. (2018). Autobin: A predictive approach towards automatic binning using data splitting. South African Statistical Journal, 52(2), 139-155.

[17] Nguyen, H. V., Müller, E., Vreeken, J., & Böhm, K. (2014). Unsupervised interaction-preserving discretization of multivariate data. Data Mining and Knowledge Discovery, 28, 1366-1397

[18] Gaughan, P. H. (2015). The International Diversification of Professional Service Firms: The Case of U.S. Law Firms [Doctoral dissertation, Cleveland State University]. OhioLINK Electronic Theses and Dissertations Center. http://rave.ohiolink.edu/etdc/view?acc_num=csu1431259487

[19] Wu, R-S, Chou, P-H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. Electronic Commerce Research and Applications. 10(3): 331-341.

[20] Noor, U., Daud, A., and Manzoor, A. (2013). Latent Dirichlet Allocation based Semantic Clustering of Heterodenous Deep Web Sources. 5th International Conference on Intelligent Networking and Collaborative Systems. IEEE, 2013.

[21] American Bar Association (1908). Canons of Professional Ethics. Available at: http://www.minnesotalegalhistoryproject.org/assets/ABA Canons (1908).pdf

[22] American Bar Association Model Rules of Professional Conduct (2023), Rule 8.4(d), Maintaining The Integrity Of The Profession.

[23] Daniels, S. and Martin, J. (1999). It's Darwinism – Survival of the Fittest: How Markets and Reputations Shape the Ways in Which Plaintiff's Lawyers Obtain Clients. Law & Policy. 21: 377-400.

[24] Zeithaml, Valarie A. (1981). How consumer evaluation processes differ between goods and services. Marketing of services 9(1): 25-32.

[25] Von Nordenflycht, A. (2010). What is a professional Service Firm? Knowledge-Intensive Firms. Academy of Management Review, 35(1): 155–174.
[26] Chaserant, C., & Harnay, S. (2013). The regulation of quality in the market for legal services: Taking the heterogeneity of legal services seriously. The European journal of comparative economics, 10(2): 267 – 291.
[27] Weigelt, K. and Camerer, C. (1988). Reputation and Corporate Strategy: A Review of Recent Theory and Applications. Strategic Management Journal. 9: 443-454.
[28] Galanter, M. and Palay, T (1990). Why the big get bigger: The Promotion-to-partner tournament and the growth of large law firms. Virginia Law Review. 747-811
[29] Fombrun, C. and Shanley, M. (1990). What's in a name? Reputation building an corporate strategy. Academy of Management Journal. 32(2): 233-258.
[30] Javalgi, R.G. and Moberg, C.R. (1997). Service loyalty: implications for service providers. The Journal of Services Marketing. 11(3): 165-179.
[31] Mishina, Y., Block, E.S., Mannor, M.J. (2012). The Path Dependence of Organizational Reputation: How Social Judgment Influences Assessments of Capability and Character. Strat. Mgmt. J. 33: 459-477.
[32] Arbel, Y. A. (2019). Reputation Failure: The Limits of Market Discipline in Consumer Markets. Wake Forest L. Rev., 54: 1239-1304.
[33] Negro, G., Hannan, M.T. and Fassiotto, M. (2015). Category Signalling and Reputation, Organizational Science. 26(2):584-600.
[34] Kherwa, P., Bansal, P. (2021). A Comparative Empirical Evaluation of Topic Modeling Techniques. In: Gupta, D., Khanna, A., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A. (eds) International Conference on Innovative Computing and Communications. Advances in Intelligent Systems and Computing, vol 1166. Springer, Singapore. https://doi.org/10.1007/978-981-15-5148-2_26.
[35] Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.
[36] Chauhan, Uttam, and Apurva Shah (2021). Topic modeling using latent Dirichlet allocation: A survey. ACM Computing Surveys (CSUR) 54(7): 1-35.
[37] Yan, X., Guo, J., Lan, Y. and Cheng, X. (2013). A Biterm Topic Model for Short Texts. In Proceedings of the 22nd international conference on World Wide Web (1445-1456).
[38] Asmussen, C.B. and Moller, C., (2019). Smart Literature review: a practical topic modelling approach to exploratory literature review. Journal of Big Data. 6(93): 1-18.
[39] American Bar Association (2004). The Lawyer Statistical Report.
[40] Rehurek R, (2009 updated Dec. 2022). Models.ldamodel – Latent Dirichlet Allocation. . Available at https://radimrehurek.com/gensim/models/ldamodel.html.
[41] Terpilovskii, M. (2023). Bitermplus 0.7.0. https://pypi.org/project/bitermplus/
[42] Basov, N., Breiger, R., Hellsten, I. (2020). Socio-semantic and other dualities. Poetics. 78:101433.

AUTHORS

Patrick H. Gaughan received his J.D. in Law from the University of Virginia, D.B.A. in International Business from Cleveland State University, M.B.A. from Trinity College (Dublin), and B.A. from Columbia University. Currently, he is an Associate Professor of Law and Assistant Dean of Global Engagement at the University of Akron School of Law.

En Cheng received her Ph. D in Computer Science from Case Western Reserve University. Currently, she is an Associate Professor of Computer Science at The University of Akron.

Taylor C. Burgess received her Masters of Science in Computer Science and Bachelor of Science in Computer Science from the University of Akron. Currently, she is working as a Software Engineer II at JP Morgan Chase.

Aine C. Bolton is completing her B.S. in Computer Information Systems from the University of Akron with intentions to pursue a M.S. in Computer Science following graduation. Currently, she is a Technology Assistant and a full-time student at the University of Akron