Bending the Law

G. Leibon[1,2], M. Livermore[3], R. Harder[2], A. Riddell[2], and D. N. Rockmore[2,4,5]

[1]Coherent Path, Inc.
[2]Department of Mathematics, Dartmouth College, Hanover, NH 03755
[3]School of Law, University of Virginia, Charlottesville, VA 22904
[4]Department of Computer Science, Dartmouth College, Hanover, NH 03755
[5]The Santa Fe Institute, Santa Fe, NM 87501

May 1, 2016

Abstract

Legal reasoning requires identification, through search, of authoritative legal texts (such as statutes, constitutions, or prior judicial decisions) that apply to a given legal question. In this paper we model the concept of the law search as an organizing principle in the evolution of the corpus of legal texts, apply that model to U.S. Supreme Court opinions. We examine the underlying navigable geometric and topological structure of the Supreme Court opinion corpus (the "opinion landscape") and quantify and study its dynamic evolution. We realize the legal document corpus as a geometric network in which nodes are legal texts connected in a weighted and interleaved fashion according to both semantic similarity and citation connection. This network representation derives from a stylized generative process that models human-executed search via a probabilistic agent that navigates between cases according to these legally relevant features. The network model and (parametrized) probabilistic search behavior give rise to a PageRank-style ranking of the texts – already implemented in a pilot version on a publicly accessible website – that can be compared to search results produced by human researchers. The search model also gives rise to a natural geometry through which we can measure change in the network. This enables us to then measure the ways in which new judicial decisions affect the topography of the network and its future evolution. While we deploy it here on the U.S. Supreme Court opinion corpus, there are obvious extensions to larger bodies of evolving bodies of legal text (or text corpora in general). The model is a proxy for the way in which new opinions influence the search behavior of litigants and judges and thus affect the law. This type of legal search effect is a new legal consequence of research practice that has not been previously identified in jurisprudential thought and has never before been subject to empirical analysis. We quantitatively estimate the extent of this effect and find significant relationships between search-related network structures and propensity of future citation. This finding indicates that influence on search is a pathway through which judicial decisions can affect future legal development.

1 Introduction

Judicial decision making is characterized by the application by courts of authoritative rules to the stylized presentation of disputed claims between competing litigants. These authoritative rules are set forth in legal source materials such as constitutions, statutes, and decisions in prior cases. For a legal source to have bearing on a current dispute, it must be retrievable by the relevant legal actors. The problem of organizing legal texts into a comprehensible whole has been recognized since Justinian I's Corpus Juris Civilis issued in 529-34. The acute problems of identifying relevant legal sources (i.e., legal precedent) presented by the common law tradition has spurred codification and classification efforts that have ranged from Blackstone's "Commentaries on the Laws of England (1765-69)" to the codification movement in the late nineteenth century [1], to the development and spread of the West American Digest System in the twentieth century [2]. Most recently, the effect of digitization on the evolution of the law, primarily in its impact on legal research, has become a subject of inquiry (see e.g., [3, 4, 5, 6, 7, 8, 9, 10]).

In this paper we consider the textual corpus of legal sources as an evolving, geometrically defined *landscape* that encompasses *regions* of the law and that is influenced by the dynamics and feedback of *law search*. Everything devolves from a model of the process of legal research in which "actors" start from a case or opinion and then build out an understanding of the relevant issues by following citations, searching for cases that cite the initial case of interest, and identifying textually similar cases. This has a natural network formulation, in which legal sources are connected to each other based on citation information and a "topic model" representation of their textual content. Topic models represent texts (embodied as a "bag-of-words") as mixtures of "topics", probability distributions over the vocabulary in the corpus (see e.g., [11]). By encoding three kinds of connectivity this becomes a *multi-network* representation, a combinatorial structure that has proved useful in a number of different contexts, such as biology and economics (e.g., [12, 13, 14]). In this work we bring the multi-network concept to the novel contexts of text-mining and text search, with a specific application to judicial texts.

Distance in this landscape reflects the ease with which a human user of the legal corpus could navigate from one legal source to another, based on the underlying citation network as well as via topical similarity, which in standard resources (e.g., through a commercial database such as Lexis-Nexis) is usually reduced to a keyword search. Our proxy for keyword navigation is a similarity network enabled via a topic modeling of the corpus. The underlying distance (metric) produces well-defined regions (i.e., groups of legal sources) that are relatively close to each other, but relatively distant from other regions. Distance is also a proxy for relevance. When new judicial decisions are issued and incorporated into the legal corpus, they interact with search technology to change the legal sources that will be discovered during the next search. This is a new kind of legal effect that, as far as we know, has never been identified as a theoretical possibility, much less formalized and subjected to an empirical test.

Use of the citation network to measure the influence of judicial opinions is now well-studied (see e.g., [15, 16, 17]), although interesting potential avenues of investigation remain underexplored (see e.g. [18] for

a citation network analysis in the context of scientific articles). On the other hand, topic models have only very recently entered legal studies where they have thus far showed promise as a new quantitative analysis framework [19, 20, 21, 22]. Both citation networks and topic modeling are examples of computational approaches to legal studies. Early conversations concerning law and digitization focused on distinction in "context" between digital and physical forms, for example, whether digitization enhanced or reduced reading comprehension or facilitated or undermined serendipity in conducting searches. In particular, the legal significance of the effects of various search modalities (citation-based, keyword, unstructured text) are only just becoming apparent (see e.g. [23]).

The landscape (metric) structure is based on a natural Markov model derived from the multi-network representation of the legal text corpus. The Markov model in turn gives rise to a natural notion of *curvature* for the underlying state space of the multi-network. As per the usual interpretation of this geometric notion, the more negative the curvature of a region of the legal landscape, the easier it is to navigate to legal sources outside that region from legal sources that are inside of the region. Curvature may change over time as new legal sources are added to the corpus. An increase in curvature in a given region indicates increasing difficulty in navigating to legal sources outside that region from within. This has the interpretation that the region has become more isolated from the rest of the legal corpus and thus is less relevant to new decisions outside of the region. We refer to this effect as the *puddling* of a region. The opposite effect wherein curvature decreases is referred to as *drainage* of a region. Drainage of a region is characterized by ease of navigation from points inside the region to legal sources that are outside the region. Notions of network curvature have only just begun to make their way into applied literature. Some early work has adapted the idea of Ricci curvature to the network setting, mainly for its relation to various isoperimetric inequalities (see e.g., [24, 25]). More recent work approaches the idea from the point of view of *optimal transport* [26]. This in turn makes strong connections to discrete Markov chains – as does ours – but this other work is quite different from the approach taken herein.

We apply our framework to an analysis of all U.S. Supreme Court cases from 1951 to 2002 and investigate the temporal evolution of the curvature over time. Key to our analysis is controlling for something we call *legal momentum* which captures that fact that some regions of the law remain active and relevant to new decisions while others become vestigial over time. A variety of social and cultural factors may explain the tendency of areas of the law to become vestigial, but, regardless of the factors involved, it is useful to focus on the active regions. When looking at the active part of the legal corpus, we find that regions of the law that experience puddling are less likely to be relevant to future cases, while regions of the law that experience drainage are more likely to be relevant to future cases.

2 Results

We have a data analytic result wherein we show that the metrics we have developed to determine "impact" of an opinion allow us to predict its ongoing relevance. We postpone the technical details of the construction to the next section. As indicated above, our results depend on a notion of distance derived from a multi-network built on the corpus of Supreme Court opinions. The multi-network is realized via the incorporation of three kinds of edges:

- "Cited by" edges – such an edge from a to b represents that opinion a is **cited by** opinion b.
- "Cited" edges – such an edge from a to b represents that opinion a **cites** opinion b and
- "Similarity edges" – this is a weighted edge between a and b (i.e., symmetric in a and b) that encodes a kind of textual comparison of the texts that depends on a *topic model* representation of the opinions (see Methods – Section 3 – for details).

Using these three kinds of edges we create a Markov chain on the space of opinions, which in turn gives rise to a notion of distance between any two opinions A and B, which we denote as $\mathit{PageDist}(a,b)$. The Markov chain further enables us to construct a notion of (local) *curvature* on this multi-networked set of opinions. For a state (opinion) a let $\kappa(a)$ denote the curvature at a. Like the traditional differential geometric notion of local curvature (curvature at a point), it reflects the ease of escape from the neighborhood nearby the point: the more *negative* this value, the easier it is to escape.[1]

If the degree of difficulty of escape is large, a walk will have a tendency to "get stuck" in the neighborhood of the state. This can be interpreted as an opinion that doesn't connect usefully beyond its surrounding or nearby opinions. Conversely, a more "fluid" area around an opinion suggests that it engages usefully with the broader opinion landscape. This kind of idea will be key to understanding the *impact* and *momentum* of an opinion.

As the network of opinions evolves, a measure of change in the local connectivity of the opinions can be expressed in terms of changing κ. We think of it as measuring how the network is *bending*. Suppose now that we consider the network at two different time points t_0 and t_1 with corresponding node (opinion) sets N_0 and N_1. We want to be a little careful as to how we measure the effect of the introduction of new cases and to that end we define $\kappa(a;N_0,N_1)$ to be the curvature of the induced chain obtained by lumping into a single connection all opinions that enter the corpus between times t_0 and t_1 that connect a pair of opinions. Basically, it allows for the new cases to enable potential "shortcuts" not present in the t_0 corpus. We then quantify a change in the induced exploration geometry as

$$\text{Bending}(N_1,N_0)(a) = \kappa(a;N_0,N_1) - \kappa(a;N_0)$$

[1] The classic example of point of negative curvature is the center of a saddle – a marble placed at such a point quickly slides off the saddle. The flatter the saddle, the closer to zero is the curvature at this center point.

where $\kappa(a;N_0)$ is the curvature at a as a point in the multi-network built on N_0 (i.e., at time t_0). Identifying the network with the timestamp we might also write

$$\text{Bending}(a;t_1 > t_0) = \kappa(a;t_1 > t_0) - \kappa(a;t_0).$$

Figure 1 shows the distribution of $\kappa(*;1990)$ as well as bending relative to 1995 in the Supreme Court opinion corpus ($Bending(*;1995 > 1990)$).

Bending is easy to interpret, it indicates whether the induced geometry at a point evolves in such a way that it became easier or more difficult to escape from the point. Regions where it is more difficult to make such transitions we call *puddling regions* and regions where it is easier are called *drainage regions*. A precise definition should work with the distribution of Bending values, so we call the subset corresponding to the bottom quartile of $\text{Bending}(*;t_1,t_0)$ the *Drainage* region (relative to a given date) – or $Drainage(t_1,t_0)$. Similarly, we call the subset corresponding to the top quartile of $\text{Bending}(*;t_1,t_0)$ the *Puddling* region (relative to a given date) – or $Puddling(t_1,t_0)$.

To make precise the utility of these definitions we first quantify what it means for a case to be "impactful". For this, we keep the notation of N_t as the set of nodes (opinions) at time t. Given $t_2 \geq t_1 \geq t_0$, define the *set of impactful cases* (at some threshold d) as

$$\text{Impact}_{t_2,t_1,t_0,d} = \{a \in N_{t_0} \mid \text{PageDist}(a,b) < d, \text{for some b } \in N_{t_2} - N_{t_1}.\}$$

Thus, this set (with these parameter values) comprises the "early" opinions a (i.e., those that could serve as precedent) that find themselves close to newly arrived (later) opinions (those issued in the period between t_1 and t_2). Thus the opinions in $\text{Impact}_{t_2,t_1,t_0,d}$ have remained relevant to the new opinions.

The threshold d can be set based on various criteria. A natural way to set it is by taking into account the PageDist distribution. A guiding principle that we often follow sets d according to the percentage of cases that we want to declare as "impactful" over a given initial or baseline period. That is, for fixed time periods $t_0 < t_1$, as d increases, so does the fraction of opinions in the corpus at time t_0 that are considered impactful. Conversely, as the fraction of cases that will be viewed as impactful grows, this implicitly corresponds to an increased threshold d.

We further define the *Initial Impact Probability (IIP)* (for $t_1 > t_0$ and a given threshold d) as the fraction of opinions present at time t_0 that are in $\text{Impact}_{t_1,t_0,t_0,d}$ – i.e., those opinions that remain impactful at time t_1 according to a threshold d. The goal is to understand how to predict which cases remain impactful as time goes on. Figure 2 shows how IIP varies with the impact on future cases $P(x \in \text{Impact}_{t_2,t_1,t_0,d} \mid \text{Impact}_{t_1,t_0,t_0,d})$. Therein we graph

$$[P(x \in \text{Impact}_{t_2,t_1,t_0,d} \mid \text{Impact}_{t_1,t_0,t_0,d}) - \text{IIP}]$$

(with $t_0 = 1990$, $t_1 = 1995$, and $t_2 = 2000$) against IIP (recall that as d increases monotonically with IIP, so that we can view both axes as functions of d). This behaves as might be expected, with an increasing

percentage of opinions remaining impactful, until such a time as too many initial cases are tossed in, some of which will be opinions that have become vestigial.

Let us now fix d so as to correspond to the maximum IIP setting in Figure 2. With the choice of d set, we now have fixed the parameter by which we identify opinions as impactful. We can now examine how drainage and puddling effects the impact on future cases. This is shown in Figure 3. We see the impact on future cases (the blue line) compared to impact on future cases in the "drainage" and "puddling" regions. Therein we see that indeed, drainage regions (low bending) have roughly a greater than 10% chance of having an impact on future cases than do puddling regions (high bending). That is, the drainage regions that are connecting up the space are more associated to future impact. The caption for Figure 3 contains some detail around the statistical significance of the result.

3 The Mathematical Framework

3.1 A random walk model for legal research

The geometry we construct for the legal corpus is based on a model of how the legal corpus is utilized as a network – that is, the geometry is derived from a model of the search process. We frame legal search as a process of "local" exploration of the opinion corpus, i.e., modeling the way in which a user of the legal corpus might navigate from opinion to opinion in the process of researching an issue. This kind of navigation is naturally viewed as a *Markov chain* (see e.g., [27]), formulated as a matrix T of *transition probabilities* where the entries are indexed by the opinions. For opinions a and b the value of the entry $T(a,b)$ is the probability of "moving to" opinion b from opinion a in an exploration of the legal corpus. More precisely, framing this as a "random walk" in "opinion space" this is the probability of moving at the next step to case b, given that you are at case a, i.e., the *conditional probability*

$$T(a,b) = P(b|a),$$

in standard notation.

Our transition probabilities are constructed as a combination of a several terms, reflecting a model of navigation of the space of legal opinions.[2] One of the defining features of judicial opinions is their citation of relevant prior legal sources.

Our model of legal search thus makes use of a combination of **three** basic types of local exploration from an initial opinion a: consideration of (1) opinions cited by a; (2) opinions that cite a, and (3) opinions that are *similar* to a from a textual point of view. The last of these is to be determined by a notion of similarity

[2]Other legal sources, including statutes and constitutions, have other types of internal ordering (such as organization by chapter or article) that may be relevant for law search. For purposes of this analysis, we restrict our application to the body of U.S. Supreme Court decisions and do not incorporate other sources of law. The framework of search that we develop, however, is generalizable to these other legal sources.

based on the use of a *topic model*. The topics are derived automatically from the overall corpus (see [11] for a friendly explanation of topic modeling). While there are a number of different kinds of topic models, the "latent Dirichlet allocation" (LDA) model (the "Dirichlet" refers to an underlying assumption of a Dirichlet distribution in the model) is perhaps the best known and most widely used [28]. This is the topic model that we use here. A detailed description of topic modeling is beyond the scope of this paper. Suffice to say that a topic model derives a representation of each text in the corpus as a mixture of probability distributions over the vocabulary in the corpus. Each distribution is a "topic".

As mentioned, the Markov chain (transition matrix) can be written as a linear combination of chains, $T_{\text{cited}}, T_{\text{cited-by}}$, and T_{sim}. Moreover, it is possible that the exploratory mode (i.e. the weights given to the three forms of connection in the network) may vary for a given search. That is,

$$T(a,b) = p_{\text{cited}}(a)T_{\text{cited}}(a,b) + p_{\text{cited-by}}(a)T_{\text{cited-by}}(a,b) + p_{\text{sim}}(a)T_{\text{sim}}(a,b) \tag{1}$$

with the proviso that

$$p_{\text{cited}}(a) + p_{\text{cited-by}}(a) + p_{\text{sim}}(a) = 1$$

reflecting that these are all the possible ways in which one navigates the corpus. (The notation suggests the weights may vary depending on the initial state of the search.)

The transition matrices T_{cited} and $T_{\text{cited-by}}$, based on the citation network are straightforward to construct. A natural and standard choice is to weight equally all opinions cited by a given opinion, and similarly for all opinions that cite the given opinion. This could be varied in some way, perhaps accounting for some notion of the importance of an opinion. We choose to work with equal weights. We make use of the excellent "Supreme Court Citation Network Data" database created by Fowler and Jeon [29].

The construction of T_{sim} requires more detailed explanation. We only consider as relevant to a given opinion the "top" topics and similarly for a given topic, only consider as relevant to our exploration those opinions who express it most strongly. More precisely, we fix integer parameters M_T and M_O such that for a given opinion a, Topic_a is the set of the M_T most heavily weighted topics in opinion a and for a topic t within Topic_a, we let Opinion_t comprise the M_O other opinions in which t was most strongly expressed. Thus for a given opinion a we can create an $M_T \times M_O$ matrix in which the i,j entry is the jth most significant opinion in the corpus for the ith most significant topic in opinion a.[3] If we define $W_{a,b}$ to be the number of times opinion b occurs in this matrix, then T_{sim} is the random walk produced by normalizing according to these weights.

The Markov chain that we have derived to mimic the search process is a natural generalization of the famous

[3]Notice that by assuming that cases are equally relevant a priori we have for a fixed Topic_k, $P(\text{opinion}|\text{Topic}_k) = \frac{P(\text{opinion})}{P(\text{Topic}_k)} P(\text{Topic}_k|opinion) \propto P(\text{Topic}_k|opinion)$ so we can form this ordering from our topic model as well.

PageRank algorithm [30].[4] Of interest to us is the geometry that this search model produces.[5] In particular, this kind of Markov-based search produces a metric on the network space that we call *PageDist*. We call the induced geometry an *exploration geometry*.

To define *PageDist* we attach one last parameter r to the random walk of (1): at each step assume a probability r of ending the exploration. Hence, starting at an opinion a the expected time (number of steps it takes) for a search to end at opinion b is

$$R(a,b) = \sum_{k=0}^{\infty} r^k T^k \Big|_{(a,b)}.$$

With this we define the *PageDist* metric as

$$\text{PageDist}(a,b) = ||R(a,\cdot) - R(b,\cdot)||_p$$

where p denotes the p-norm.[6] The *PageDist* metric captures our notion of distance within the landscape. Figure 4 shows the distribution of distances among our corpus of Supreme Court opinions.

The random walk setting also makes possible a definition of *curvature* that encodes a level of difficulty for escape from a given point (in the execution of a random walk). If the degree of difficulty is large, a walk will have a tendency to get "stuck" in the neighborhood of the state. This can be interpreted as an opinion that doesn't connect usefully with its surrounding or nearby opinions. Conversely, a more "fluid" area around an opinion suggests that it engages usefully with the broader opinion landscape. This kind of idea will be key to understanding the *impact* and *momentum* of an opinion. We define curvature as

$$\kappa(a) = \log(R(a,a) - 1).$$

As the network evolves we measure how local connectivity in terms of changing κ. We think of it as measuring how the network is *bending*. Let us make this precise. Given a network N with a transition matrix P reflecting a Markov process on the network, let $S < N$, be node sets. A Markov chain on N induces a chain on S by using the weights

$$W_S(a,b) = P(a,b) + \sum_{k \in N \setminus S,\ a \neq b} P(a,k)P(k,b),$$

for $a,b \in S$. Note that we are simply lumping together into one term all transitions a to b that go outside of S. We form a new transition matrix $P(a,b;S,N)$ normalizing $W_S(a,b)$ so that the weights sum to one at each vertex. We call this the *induced local exploration*. This induces a corresponding exploration geometry and a curvature κ for S relative to N which we denote as $\kappa(a;S,N)$. This is the curvature notion introduced above.

[4] It is worth noting that another natural candidate for a textual geometry includes [31] wherein the concept of a *network with directions* is introduced. Therein, "directions" function as "points at infinity", producing a hyperbolic metric on the network. For this – and any text corpus – the pure topics provide an obvious choice of direction.

[5] We are indebted to Peter Doyle for early conversations regarding the geometrization of Markov chains and PageDist.

[6] Recall that this notation means $(\sum_x [R(a,x) - R(b,x)]^p)^{1/p}$.

As per the description in Section 2 we consider the network at two different time points t_0 and t_1 with corresponding node sets N_0 and N_1. Then we can quantify a change in the induced exploration geometry as

$$\text{Bending}(N_1, N_0)(a) = \kappa(a; N_0, N_1) - \kappa(a; N_0, N_0).$$

Identifying the network with the timestamp we might also write

$$\text{Bending}(a; t_1 > t_0) = \kappa(a; t_1 > t_0) - \kappa(a; t_0).$$

4 Closing thoughts

In this paper we introduce a new multi-network framework integrating citation and textual information for encoding relationships between federal judicial opinions. The citation component derives from the underlying citation network of opinions. The textual piece derives from an LDA topic model computed from the text corpus. The network is the reification of a basic model of legal search as would be executed by a prototypical legal researcher ("homo legalus") looking for cases relevant to some initial case. The notion of search turns into a Markov chain on the network, built as a linear combination of the individual chains on the citation and topic networks. The Markov process produces a notion of *distance* between opinions which can also be thought of as a proxy for relevance. Along with distance, there is a notion of curvature, and with this an implicit framing of the opinion corpus as a "landscape" which we call "the legal landscape". We have implemented a first generation website that will allow users to explore a smallish subset of Supreme Court opinions using this search tool (*www.bendingthelaw.org*).

The text corpus evolves in the sense that cases enter the corpus regularly and in so doing continually deform the associated text landscape. Of particular interest are those cases that remain relevant over long periods of time – such cases we call *impactful*. Some regions of the legal landscape have the property that they serve as nexuses of connection for regions of the landscape. We show that those regions which over time become significantly more negatively curved are such connective areas. With the analogy of flow in mind, we call such areas, regions of "drainage". Areas which experience a significant increase in curvature we call "puddling regions". We show that drainage areas are more likely to contain the impactful cases than the puddling regions. We further show that opinions that start off impactful, in the sense of entering the landscape highly relevant to many cases over a short period of time tend to remain impactful, thereby suggesting a property of *legal momentum*.

There are natural next steps to take with this idea. In one direction we will expand the text corpus to include all Supreme Court and Appellate Court Opinions. We also plan to validate and compare our model by asking users to compare the results of our search algorithm (under a range of parameter choices) with their own usual research approaches. Our newly introduced opinion distance function gives a new variable to explore the relations of opinions to all kinds of social and economic variables. It is also natural to export this model to other court systems that produce English language opinions. In this regard it would be interesting

to see the ways in which the "bending" of the courts systems vary, and try to understand what might account for such (possible) variation. Ultimately, it would also be of interest to effect the integration of distinct corpora via this model.

References

[1] Garoupa N, Morriss AP (2012) The fable of the codes: The efficiency of the common law, legal origins and codification movements. University of Illinois L Rev 5: 1443.

[2] West JB (1909) Multiplicity of reports 2. Law Library Journal 4.

[3] Katsh E (1993) Law in a digital world: Computer networks and cyberspace. Vill L Rev 38: 403.

[4] Berring RC (1986) Full-text databases and legal research: Backing into the future. Berkeley Tech L J 1: 27.

[5] Berring RC (1987) Legal research and legal concepts: Where form molds substance. Cal L Rev 75: 15.

[6] Hanson FA, Allan F (2002) From key numbers to keywords: How automation has transformed the law. L Lib J 94: 563.

[7] McGinnis JO, Wasick S (2014) Law's algorithm. Fl L Rev 66: 991.

[8] Hellyer P (1997) Legal positivism as legal information. Cornell L Rev 82: 108.

[9] Eckmann JP, Moses E (2005) Assessing the influence of computer-assisted legal research: A study of california supreme court opinions. L Lib J 97: 285.

[10] Fronk CR (2010) The cost of judicial citation: An empirical investigation of citation practices in the federal appellate courts. U Ill JL Tech & Policy 2010: 5825–5829.

[11] Blei DM (2012) Probabilistic topic models. Communications of the ACM 55: 77-84.

[12] Barigozzi M, Fagiolo G, Mangioni G (2011) Identifying the community structure of the international-trade multi-network. Physica A: Statistical Mechanics and its Applications 390: 2051–2066.

[13] Blinov ML, Udayar A, Yarbrough W, Wang J, Estrada L, et al. (2012) Multi-network modeling of cancer cell states. Biophysical J 102.

[14] et al MK (2014) Multilayer networks. Journal of Complex Networks .

[15] Fowler JH, Johnson TR, Spriggs I, James F, Jeon S, et al. (2007) Network analysis and the law: Measuring the legal importance of supreme court precedents. Political Analysis 15: 324-346.

[16] Fowler JH, Jeon S (2008) The authority of supreme court precedent. Social Networks 30: 16-30.

[17] II MJB, Katz DM, Zelner J (2009) Law as a seamless web? comparison of various network representations of the united states supreme court corpus (1791-2005). In: Proceedings of the 12th International Conference on Artificial Intelligence and Law (ICAIL 2009).

[18] Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. Science 342: 468–472.

[19] Rice D (2012) Measuring the issue content of supreme court opinions through probabilistic topic models. Presentation at the 2012 Midwest Political Science Association Conference, Chicago, Illinois .

[20] Nardi DJ, Moe L (2014) Understanding the myanmar supreme court's docket. In: Crouch M, Lindsey T, editors, Law, Society and Transition in Myanmar.

[21] George C, Puri S, Wang DZ, Wilson JN, Hamilton W (2014). Smart electronic legal discovery via topic modeling.

[22] Livermore M, Riddell A, Rockmore D (2015) A topic model approach to studying agenda formation for the u.s. supreme court. Virginia Law and Economics Research Paper No 2015-2 .

[23] McGinnis JO, Wasick S (2014) Laws algorithm. Fl L Rev 66: 991.

[24] Chung F, Yau ST (1996) Logarithmic harnack inequalities. Math Res Lett 3: 793-812.

[25] Lin Y, Yau ST (2010) Ricci curvature and eigenvalue estimate on locally finite graphs. Math Res Lett 17: 345-358.

[26] Ollivier Y (2009) Ricci curvature of markov chains on metric spaces. Journal of Functional Analysis 256: 810-864.

[27] Grinstead CM, Snell JL (1997) Introduction to Probability. American Mathematical Society.

[28] Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. Journal of Machine Learning Research 3: 993–1022.

[29] Supreme Court Citation Network Data. URL `http://jhfowler.ucsd.edu/judicial.htm`. Accessed January, 2015.

[30] Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. In: Crouch M, Lindsey T, editors, Computer Networks And ISDN Systems, Elsevier. pp. 107-117.

[31] Leibon G, Rockmore DN (2013) Orienteering in knowledge spaces: The hyperbolic geometry of wikipedia mathematics. PLoS ONE .

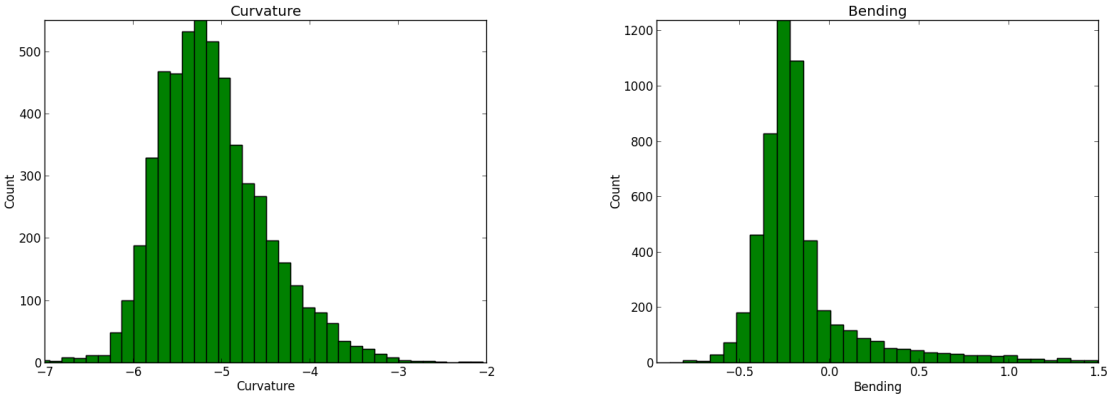

Figure 1: On the left we see a histogram of the the curvature $\kappa(*;1990)$ and on the right we see the bending $\text{Bending}(*;1990,1995)$. This gives a sense of the variation of the curvature over time.

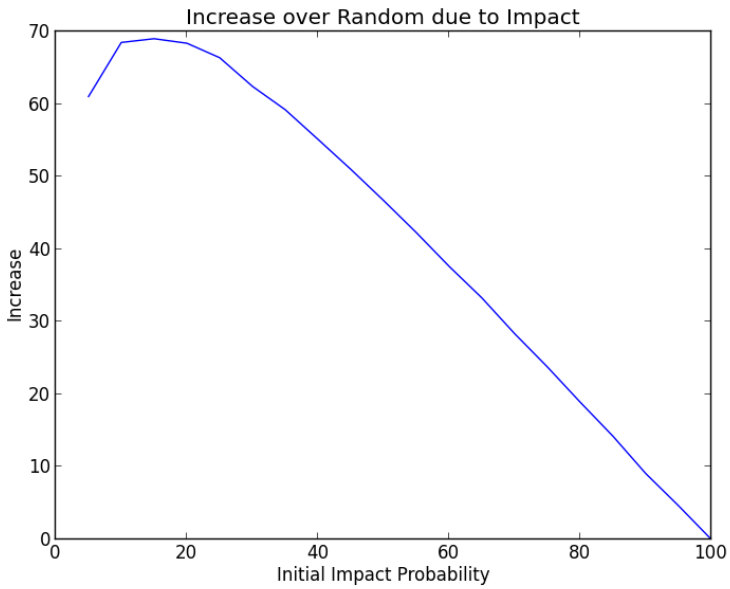

Figure 2: Here the x-axis is Initial Impact Probability (as a percentage), the function of d that gives the fraction of early opinions with impact on the initial set of new opinions. Recall that as IIP increases, so does d. In blue we see we see $[P(x \in \text{Impact}_{t_2,t_1,t_0,d} \mid \text{Impact}_{t_1,t_0,t_0,d}) - \text{IIP}]$ with $t_0 = 1990$, $t_1 = 1995$, and $t_2 = 2000$ (and d a function of IIP). Thus, this is the proportion of early (pre-1990) opinions that continue to have impact in the 1995-2000 period, given that they had impact in the 1990-1995 period, minus the fraction of opinions that initially have impact on opinions written between 1990 and 1995. Thus, we are subtracting out some baseline guess of how many of these early cases you would expect to have impact in this time based on earlier information. This measures how much larger than random the future impact is given recent impact. This is all a function of d or equivalently, IIP. We see that IIP $= 20$ is roughly an optimal value.

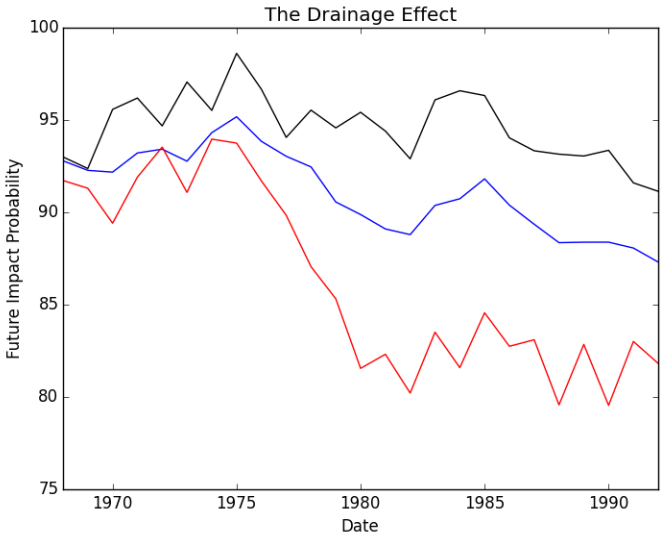

Figure 3: Here the x-axis is the year the case was decided. The blue curve is $P(x \in \text{Impact}_{t_2,t_1,t_0,d} \mid \text{Impact}_{t_1,t_0,t_0,d})$ with $t_0 = date$, $t_1 = date + 5$, and $t_2 = date + 10$ and d fixed by the 20th percentile, as in Figure 2. In black we see the same curve conditioned on *Drainage* regions, while in red the same curve conditioned on *Puddling* regions. Notice that indeed, the bending is correlated with long term impact as predicted, and that after the geometry has really "warmed up" (about 1978), we see a fairly stable 10% difference. To confirm that this correlation is statistically significant, let the null hypothesis be that there is nothing but a random difference between the *Drainage* and *Puddling* regions. So for a fixed measurement, under the null hypothesis there would be a fifty-fifty chance that we confirm our suspicion (technically, bounded by 50% when allowing for ties). Furthermore, for events that differ by at least 5 years, the $N_{t_2} \backslash N_{t_1}$ populations are distinct , so that the measurements are suitably independent. Thus, we have 6 independent measurements with a perfect track record and can conclude that the p_{value} is less than $\frac{1}{2^6}$ and the correlation significant.

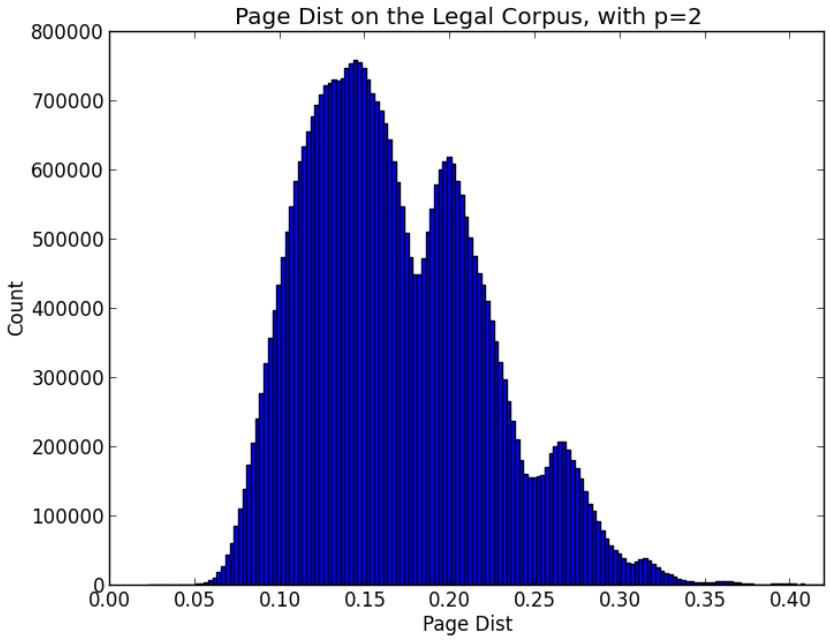

Figure 4: Here we see a histogram of the PageDist values when computed on the legal corpus. We choose $p = 2$, $r = \frac{1}{2}$, and $M_T = M_C = 10$.

5 Appendix: Implementation.

The ideas presented in this paper form the foundation of new web-based search tool for exploring a space of legal decisions using the exploration geometry introduced in the body of this paper. Specifically, we have built a prototype for a website and user interface that will enable the exploration of an opinion database, that ultimately will encompass all Federal Court and Supreme Court cases. At present it is running on a small subset (SC cases 1950–2001). This prototype can be found at *www.bendingthelaw.org*.

The current user interface (UI) introduces users to cases in the "vicinity" (in the sense of our exploration geometry) of a pre-identified case specified by the user. The anticipation is that these cases will be strong candidates for precedent-based reasoning. As per (1) the return depends on the database of cases as well as the individual weights assigned to the three-component random walk process encoding the exploration geometry – that is, a choice of weights $p_{cited}, p_{citedby}$, and p_{sim}. As a first step we allow a choice of weights from $\{0,1,2\}$. Recall that the similarity piece of the random walk, T_{sim} requires that we construct the "topic by opinion" matrix of a given size. We choose that to be 10 – i.e., that for any given topic we consider the 10 opinions that make the most use of it and conversely, for any opinion, we consider the 10 topics that make the strongest contribution to it.

Given an initial query, the UI provides two complementary representations: (1) a ranked list of geometrically closest (in terms of PageDist) cases and (2) a map of the space, centered on a case of origin (the original input). As a "map", this representation shows not only the relation of cases to the initial query, but also the relations of the closest cases to each other. The associated visual integrates the citation network representation with a 2-d multidimensional scaling visualization of the thirty (including the query) intercase distances. (An arrow to from case A to case B means that case A cites case B.) The map is generated by clicking on "View Case Network" (after executing the query). The opinion map produced from the query "329 US 187: Ballard v. United States" is shown in Figure 5.

Figure 5: Here is a snapshot from our alpha version UI for exploring the space of legal opinions. The current UI is built on the database of Supreme Court opinions over the time period 1950–2001. What we see here is the 2-d MDS visualization of the PageDist neighborhood of 30 closest cases to "329 US 187: Ballard v. United States". Note that the exploration weights have been set to 2 ("cited"), 1 ("cited by"), and 2 ("topic similarity").