frontiers in Physics

ORIGINAL RESEARCH
published: 31 March 2021
doi: 10.3389/fphy.2020.625241

A Multinetwork and Machine Learning Examination of Structure and Content in the United States Code

Keith Carlson[1], *Faraz Dadgostari*[2], *Michael A. Livermore*[3]* *and Daniel N. Rockmore*[1,4]

[1]*Department of Computer Science, Dartmouth College, Hanover, NH, United States,* [2]*Department of Systems Engineering, University of Virginia, Charlottesville, VA, United States,* [3]*School of Law, University of Virginia, Charlottesville, VA, United States,* [4]*The Santa Fe Institute, Santa Fe, NM, United States*

This paper introduces a novel linked structure-content representation of federal statutory law in the United States and analyzes and quantifies its structure using tools and concepts drawn from network analysis and complexity studies. The organizational component of our representation is based on the explicit hierarchical organization within the United States Code (USC) as well an embedded cross-reference citation network. We couple this structure with a layer of content-based similarity derived from the application of a "topic model" to the USC. The resulting representation is the first that explicitly models the USC as a "multinetwork" or "multilayered network" incorporating hierarchical structure, cross-references, and content. We report several novel descriptive statistics of this multinetwork. These include the results of this first application of the machine learning technique of topic modeling to the USC as well as multiple measures articulating the relationships between the organizational and content network layers. We find a high degree of assortativity of "titles" (the highest level hierarchy within the USC) with related topics. We also present a link prediction task and show that machine learning techniques are able to recover information about structure from content. Success in this prediction task has a natural interpretation as indicating a form of mutual information. We connect the relational findings between organization and content to a measure of "ease of search" in this large hyperlinked document that has implications for the ways in which the structure of the USC supports (or doesn't support) broad useful access to the law. The measures developed in this paper have the potential to enable comparative work in the study of statutory networks that ranges across time and geography.

Keywords: multinetwork, statutory network, United States Code, topic modeling, assortativity, law search

OPEN ACCESS

Edited by:
Daniel Martin Katz,
Illinois Institute of Technology,
United States

Reviewed by:
Satyam Mukherjee,
Indian Institute of Management
Udaipur, India
Jie Cao,
Nanjing University of Finance and
Economics, China

***Correspondence:**
Michael A. Livermore
mlivermore@virginia.edu

Specialty section:
This article was submitted to
Social Physics,
a section of the journal
Frontiers in Physics

Received: *02 November 2020*
Accepted: *30 December 2020*
Published: *31 March 2021*

Citation:
Carlson K, Dadgostari F, Livermore MA and Rockmore DN (2021) A Multinetwork and Machine Learning Examination of Structure and Content in the United States Code. Front. Phys. 8:625241. doi: 10.3389/fphy.2020.625241

1 INTRODUCTION

In this paper we present a network-based framing and analysis of the United States Code (USC), the legal corpus comprising the federal statutes of the United States. Statutes therein possess hierarchical structure via their organization into titles, sections, sub-sections and the like, and a cross-reference structure in which provisions of a code cite to other provisions for purposes of sharing definitions or establishing legal relations. These overlapping structures of hierarchy and cross-references intertwine content which can also be studied as a network through a similarity structure derived from its defining documents. Taken together, these interleaved network structures define the USC – and more generally, any statutory network corpus – as a *mutilayered network* or *multinetwork*, a

complex network structure increasingly of interest in the areas of network science and complex systems [1]. The contribution of this work is the first framing of the USC as a multinetwork and with that, a first network-based analysis of the USC that includes the calculation of various network-related metrics (degree, betweeness centrality, hubs and authorities measures) and other quantifiable characteristics of the USC as well as detailing the relationships between the layers. In doing so, we present a defining framework for the notion of a "statutory network" and a collection of attendant measures that will enable future work in comparative and intrinsic statutory analysis.

As discrete, information dense, and legally important corpora, bodies of statutes have proven particularly attractive to researchers who are interested in applying computational tools to legal texts [2]. The hierarchical and networked structure we see in the USC is a hallmark of complex systems, which include regulatory structures as well as evolved organizational structures ranging from corporations to societies and ecosystems (*see* e.g., [3, 4]).

As mentioned, network characteristics, as well as the large body of work in quantitative textual analysis of document corpora enable an empirical approach to the study of statutory networks. This puts the study of statutes squarely within the vibrant body of research that is bringing to bear the tools of complex systems and machine learning on legal documents and the law (*see* [5] for a recent survey). For example, it may be that there are temporal or geographic determinants of a statutory network that may influence the diffusion of legal culture over space and time [6]. There may also be consequences of certain statutory network features, such as regulatory complexity, that increase legal transaction costs [7] or lead to hidden "cumulative" costs of regulation [8]. In this vein, new measures of statutory complexity are needed to move the field beyond fairly primitive proxies, such as word counts or simple *n*-gram style metrics (see e.g., [9, 10], as well as [11] for analogous efforts in the banking industry).

Network features of the law can also be exploited to study certain types of legal behavior. Prominent in this regard is the body of work that has been done related to the citation network of Supreme Court opinions, notably the groundbreaking work of Fowler and his collaborators in their study of precedent [12]. The use of text analysis tools in the law is more recent. Early uses of the machine learning *topic modeling* approach used herein include novel analyses of the impact of Supreme Court opinions on litigation [13], the influence of clerks on opinion-writing [14] and a more general study of opinions as genre [5]. Extensions to the much larger corpus of Circuit Court opinions has resulted in a new revelation of publication bias in that setting [15].

The multinetwork framework of SCOTUS opinions combining citations with topic-similarity has further served to produce a geometric framework for their study [16] and is used to study the problem of "law search" [17]. This work provides useful insights into an important category of legal behavior that has been difficult to study using traditional tools. Indeed, as far back as Jeremy Bentham, legal philosophers have recognized that the diffuse nature of the law (especially in common law systems) poses important normative problems [18]. Given the growth of publicly available legal datasets[1], advances in understanding the internal organization of the law and using that understanding to facilitate search of legal materials for non-experts and experts alike may also facilitate broader access to the law in a comprehensible format. These questions are similarly germane to the USC.

This paper is the first to merge information on statutory structure with the semantic content of statutory text. We thus build and expand on prior work, such as [2, 19]; that focuses on organizational features (specifically hierarchical structure and cross-references) of statutory codes. The inclusion of semantic content is particularly important: while the hierarchical and cross-reference structure of a statutory corpus provides some valuable information about the nature of a legal regime, the structure itself has no legal effect. Statutory structure establishes relationships and order, but it is the semantic content that ultimately is the legal *materiel* stitched together through structure. A better understanding of the relationship between statutory structure and content may also prove particularly useful in developing better search tools for statutory law. Because statutory language is relatively parsimonious—especially compared to the lengthy narrative documents produced by courts—common search approaches (including Boolean searches and various natural language augmentations) can be ineffective at identifying relevant statutory authority for a given legal matter. Systematic patterns in the overlap of structure and substance could be leveraged in search tools that were specifically designed to lower the costs of identifying relevant statutory text. While we focus on the example of the United States Code, the techniques described here could be applied to any statutory system. Empirical comparative extensions are likely to yield particularly worthwhile insights into different system-level characteristics of legal orders and the relationship of those characteristics to outcomes of interest, which could vary from sociological legitimacy to regulatory compliance costs.

For our analysis of the semantic content, we rely on the machine learning tool of *topic modeling* [20] and specifically the *structural topic model* (STM) introduced in [21]. A great deal of work has gone into the development and refinement of topic models and they have become widespread within social sciences (e.g., [22],) the humanities (e.g., [23–26],) and other text-centric disciplines. Several recent papers have applied topic models to legal documents [5, 27–29]. The STM approach builds on the conventional and widely used Latent Dirichlet Allocation (LDA) topic model.

In **Section 2** we discuss prior work and summarize the contributions of our analysis. In **Section 3** we describe our data derived from the current (online) version of the USC. This includes some basic descriptive statistics of the structural network of the USC including centrality and hub/authority measures of the underlying title network. **Section 4** contains the meat of the analysis. After providing some additional detail on

[1]See e.g., Court Listener https://www.courtlistener.com/.

topic models and their utility in producing basic characterizations of unstructured collections of documents, we report the results of our topic model of the USC as well as some intuitive descriptions of the relationships between topics and statutory structure. The topic modeling of the USC text is new and a main contribution of this work. Our next main contributions build on the results of the topic modeling to generate a set of measures to examine the relationship between semantic features of the USC and its structure, the latter represented in its cross-references and hierarchical organization into higher-level titles. We do this first by using an *assortativity* measure that connects cross-reference and semantic structure (using the topic model output). This shows significant relationship between title and content. We then construct new relational measures inspired by mutual information that investigate the degree to which semantic content can predict connectivity. Using an SVM machine learning approach, we achieve predictive accuracy of 60% in using topic proportions to predict titles. We create a second measure using the law search model developed in [16, 17] to predict cross-reference citations from topics and achieve similar performance as has been achieved for an analogous experiment using the corpus of Supreme Court opinions. Taken together, these results suggest good alignment between the structural and content layers of the multinetwork, with attendant positive implications for searchability. We close in **Section 5** with a description of future work enabled by an anticipated merging of the USC data with other legal corpora.

2 GENERAL CHARACTERISTICS OF LEGAL SYSTEMS

Statutes are the laws enacted by a legislative body. In common law jurisdictions such as the United Kingdom, statutory law (e.g., the enactments of Parliament) can be contrasted with judge-made law that accretes through the decisions of courts in individual cases. In the United States, most areas of federal and state law have statutes at their foundation, with courts charged with the task of statutory interpretation through the application of statutory language to particular cases. Administrative agencies, themselves established and empowered via statutes, frequently have a role in elucidating broad statutory commands through more detailed regulations.

Statutes are distinct from constitutions, which are adopted and altered through special procedures rather than the typical legislative process. Taking the United States as an example, Article I, **Section 7** of the U.S. Constitution sets out the procedure for Congress to adopt statutes, via majority vote in the House of Representatives and Senate and presentment to the President, and in the case of a presidential veto, a two-thirds vote in both chambers. Article V of the Constitution describes the (very difficult) procedure for amending the Constitution itself, which requires that any proposed amendment be ratified by three fourths of the states.

Statutes are a longstanding object of analysis for empirical legal study: for example, there are a substantial number of papers that examine the effects of the death penalty—a statutory provision—on crime (e.g., [30].) In addition to investigating the consequences of particular policy choices as embodied in statutes, empirical legal scholars also examine factors that affect the decision of whether or not to adopt a new law, such as prior success in another jurisdiction or geographic proximity to other adopters [31].

In addition to studying *specific* legislative enactments, scholars have also focused on certain *general* characteristics of legal systems. For example, scholars in the "legal origins" tradition have argued that certain legal characteristics that are correlated with whether a country has a common law or civil law system are associated with macro-social outcomes such as economic development [32, 33]. This literature has been broadly influential and has shaped recent political discourse on law and development [34]. For criticism of this work, see [35, 36] among others.

A related literature focuses specifically on the notion of *legal complexity* and the question of relationships between the complexity of a legal system and a variety of societal outcomes [37, 38]. At a high level of abstraction [39] argues that societal complexity along a variety of fronts eventually contributes to the disintegration of politically organized groups. More concretely, scholars have argued that legal complexity hampers economic development through several channels, including by the lowering of returns on capital and thus impeding innovation [40–43]. In recent years, several scholars have attempted to use data on aggregate regulatory levels to draw conclusions about the costs and benefits of various regulatory regimes [9, 44, 45]. In the realm of political discourse, trade associations representing regulated industry frequently bemoan legal complexity and the cumulative cost of regulations [46].

Rigorous work on legal complexity has been hampered by inadequate definition and measurement of the underlying concept (*see* e.g., [37].) Simple measures, such as the number of pages in the U.S. Federal Register, or counting n-grams that target "command" type language (i.e., "shall" or "must") have been used as rough proxies [45], but their shortfalls are fairly obvious. As domestic legal regimes cope with continued economic growth and global integration, legal complexity (broadly understood) is likely to increase, and social scientific study of this phenomenon will take on even greater importance. But that work will continue to struggle without reliable and accurate measures of the phenomenon.

A separate and related vein of scholarship examines the practice of *law search* – the process whereby agents seek out relevant legal authority to apply to a given legal problem [16, 17]. Other things being equal, a legal system in which it is more difficult to locate relevant authority can be understood as less comprehensible in way that is often attributed to its greater "complexity." In short, a legal system of greater complexity is simply one in which it is more difficult for legal actors of any level of background to learn, know, and understand their rights and responsibilities. In [17]; the authors introduce the notion of "convergence" in a legal system, which is the tendency for legal participants to *converge* via law search to a similar set of legal authorities that are relevant to a given legal question. Convergence may be inversely associated with complexity.

When possible, quantitative measures – possibly tied to the field of complexity science – have the promise of making rigorous discussions of legal complexity. In particular, when a network framework makes sense, the tools of network science can be brought to bear effectively on the subject [37] and similarly for computational text analysis [17]. Both of these approaches are germane in the case of the study of the USC and statutory networks generally. The work described in this paper contributes to the literature of legal complexity and law search, and more generally to the field of computational analysis of legal texts (and statutes specifically). The work presented herein is the first project that we are aware of that combines information on statutory structure with semantic data on legal content to study the relationship between structure and substance in a statute-based legal order. Our findings regarding the United States Code can help set the stage for comparative work that examines similar relationships within other legal systems. In addition, the techniques we use to capture semantic content and overlay content and structure (and in particular the use of topic models) can inform future efforts to define measures of legal complexity. Finally, prior efforts to study law search have focused on judicial opinions, which are particularly information-dense, and therefore relatively easy to navigate. Our work can be leveraged to expand those analyses to statutory (and regulatory) texts, where the costs of law search may be even more pronounced.

3 DATA

As per the U.S. Constitution, the fundamental role of Congress is to exercise the "legislative power," much of which is embodied in the statutes it adopts. Statutes are simply the laws adopted by Congress. These laws cover a wide range of public and private conduct – everything from the tax rate and the penalties for kidnapping to provisions establishing the authority of the Environmental Protection Agency to issue air quality standards.

When a statute is successfully adopted (e.g., via majority votes in both houses and a presidential signature), it is issued as a Public Law and published as a session law, and is compiled chronologically in the Statutes at Large.[2] These session laws are the exact text enacted by Congress, and so represent "the law" as a direct exercise of Congress's power.

These public laws can be difficult to navigate. They are not organized by subject matter. Furthermore, rather than evolving through edits, subsequent revisions are recorded as separate session laws which then operate on earlier versions. This in turn creates a complex relation of interlocking dependencies and cross-references. For several decades after the founding of the republic, the work of compiling and publishing a comprehensive representation of the current law fell to private publishers. These documents were useful for lawyers, but had no official legal status. In the 1870s, Congress undertook an official codification, the Revised Statues of the United States, which was meant to capture the existing state of the law. Subsequent efforts at official codification faltered until the USC was approved by Congress in 1926.[3] Eventually the maintenance of the USC was brought into the U. S. government and now the USC (or "the Code") is maintained and published by the U.S. Office of the Law Revision Counsel (OLRC).[4] The OLRC is also responsible for the organization of the Code and compiling relevant changes as they are enacted.[5]

The representation used in the following analyses is based on a one-time "snapshot" of the Code based on the information published by the OLRC. Data collection occurred during the period October 2016 through January 2017. For this project, we do not examine dynamic over-time effects as the USC is altered through the legislative process. Instead, we focus on static features of the Code as it existed during the data-collection period.

As will be discussed in more detail in the next section, the topic model approach we use to engage in semantic analysis relies on our treating the USC as a corpus of "documents." The quotation marks call out the distinction between a "document" in the sense of a topic model and a document in a more colloquial sense. The former is a contiguous collection of words (or even more generally, character strings extracted via a standard sort of text processing format) while the latter might suggest something with some recognizable narrative form. The USC is organized in a nested fashion with multiple levels wherein *section* is the core organizational unit, containing thematic blocks of text.[6] We treat these sections as the documents for topic modelling purposes.

The structure of the USC as a hierarchical information repository is somewhat complicated, involving different levels across the Code. These include title, (possible) subtitle, chapter, parts and subparts, subsections, paragraphs, clauses, and items. The highest (i.e., broadest) organizational unit is the "title." Titles can be thought of as a general subject heading: examples include "Armed Forces" (Title 10) and "Public Health and Welfare" (Title 42). We use only sections and titles as the relevant unit of analysis and ignore the other levels, in part because they are treated differently depending on the title. With this definition of the document unit there are 41,138 total documents in the USC organized into 44 titles.

[2]https://www.loc.gov/law/help/statutes-at-large/

[3]Pub.L. 69–441, 44 Stat. 778, enacted June 30, 1926.

[4]https://uscode.house.gov/

[5]There are some additional complications concerning how the Code is compiled from the Statutes at Large and how courts give effect to various texts, but they are not important for this study. For more background on the Code, see [52].

[6]For example, **Section 112** of Title 42 is "Removal of revenue officers from port during epidemic" and states "Whenever, by the prevalence of any contagious or epidemic disease in or near the place by law established as the port of entry for any collection district, it becomes dangerous or inconvenient for the officers of the revenue employed therein to continue the discharge of their respective offices at such port, the Secretary of the Treasury, or, in his absence, the Undersecretary of the Treasury, may direct the removal of the officers of the revenue from such port to any other more convenient place, within, or as near as may be to, such collection district. And at such place such officers may exercise the same powers, and shall be liable to the same duties, according to existing circumstances, as in the port or district established by law. Public notice of any such removal shall be given as soon as may be."

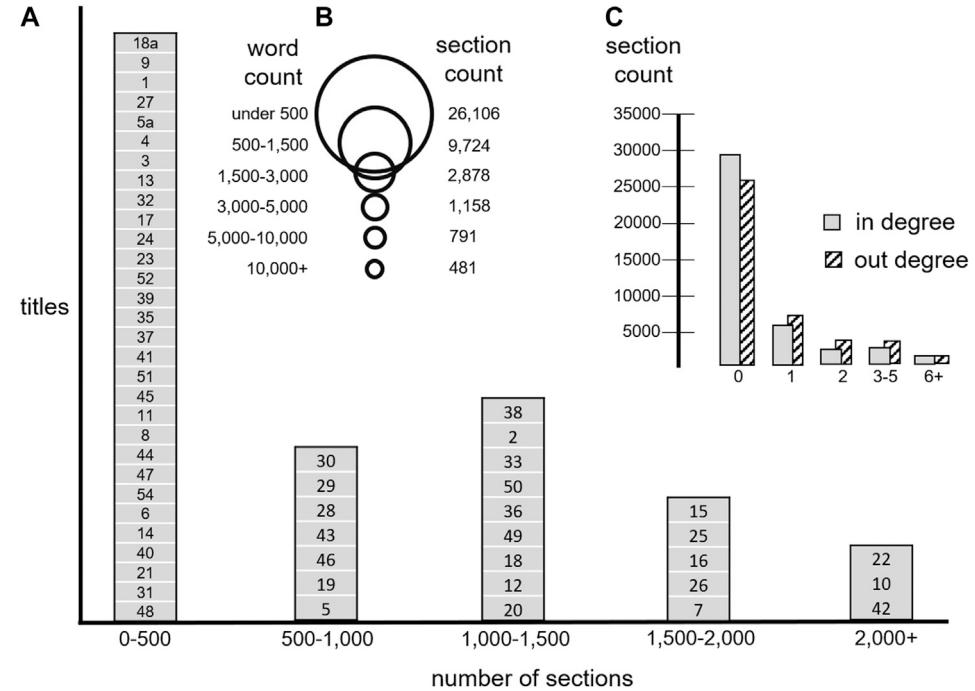

FIGURE 1 | The number of sections in each title varies considerably, and is right skewed. Panel **(A)** arranges each title from least to most sections and places the titles into four buckets. Title 18a has the smallest number of sections (17) and title 42 has the largest (4,581). There is also considerable variation in the word counts in each section, with a large number of fairly small sections (under 500 words) and a small number of very long sections (over 10,000 words). Panel **(B)** illustrates this variation. Panel **(C)** provides histograms of in-degree and out-degree for cross-references at the section level.

Figure 1 reports basic information on the word counts within sections and the counts of sections within titles.[7] There is considerable variation in the section count of the titles and the word count of the sections. Document length is right-skewed, with a modal length less than 500 words and a small number of much longer outliers. The average number of words per document is 1,036. There is a right-skewed distribution for section within titles as well, with many titles containing only a few hundred or even just several dozen sections. A second common size of around one to two thousand sections also is found. There are a few very large outliers, with Title 42 ("Public Health and Welfare") the largest.

Our data also includes information on cross-references within the Code. These section-to-section links (along with the sections) produce the (second) inherent network structure of the USC. A directed edge is created if one document (i.e., section) references another. **Figure 1** also reports histograms of the number of in- and out-citations by section, showing that most sections have neither incoming nor outgoing edges, and few have greater than six (either incoming or outgoing).

There may be multiple citations between documents, but in our representation, a single edge is constructed between two documents if there is one or more citations. We do not construct edges for citations above the section level (e.g., when a citation is to an entire title or chapter). With these caveats in place there are 38,399 cross-references between sections. Note that in the citation network, there are fewer edges than there are nodes. In the USC, there are a large number of sections that do not include cross-references to any other sections, and are not cross-referenced by any other section. This is not necessarily surprising. Many sections are self-contained and do not need to make reference to other sections for shared definitions or other purposes. Likewise, many sections provide no more general terms that must be referenced elsewhere. In addition, there are other relations that exist between statutory sections that may not be called out via cross-reference. For example, the location of a section within a chapter or other supra-section category may be meaningful and denote certain types or relationships. Some sections include intra-section cross-references. These show up as loop edges in the citation network. As will be discussed later in this paper, there is a substantial amount of overlap between the cross-reference network and the hierarchical structure of the USC, with many cross-references occurring within title. However, there is a fair amount of inter-title cross-referencing as well.

We calculate several additional network statistics on the basis of the analysis above, including graph density, average total degree, variance of in-degree and out-degree, and the ratio of the number of components over the number of nodes. They are reported in the Appendix in **Supplementary Table S1**.

[7]A list of title names is provided in the Appendix in **Table A7**.

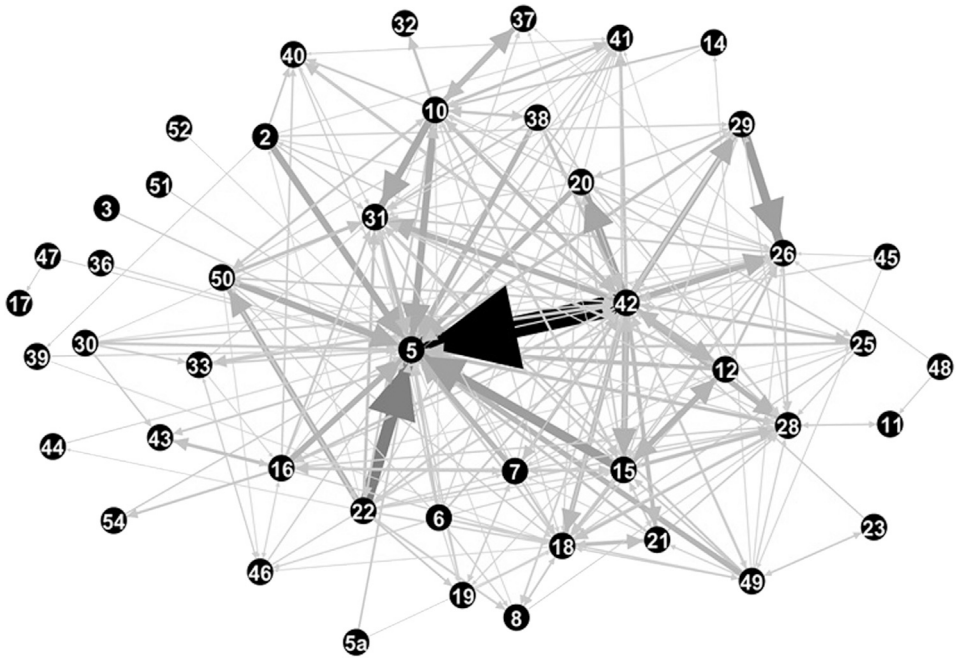

FIGURE 2 | Network representation of cross-references at the title level. Title 42 "Public Health and Welfare" has the largest number of outgoing citations, and Title 5 "Government Organization and Employees" has the most incoming citations.

Figure 2 provides an illustration of the citation network, highlighting only inter-title references. The edge thickness reflects the number of inter-title cross-references. As is visually apparent some titles play a more central role in the inter-title network than others and some titles are largely self-contained.[8] The most obvious pattern found in this figure is that Title 5, "Government Organization and Employees," has a very large number of incoming citations—presumably because it contains language or arrangements that are quite general and apply across legal categories—whereas the massive and hodgepodge Title 42 ("Public Health and Welfare") has a large number of outgoing citations, presumably because of its size and catchall nature.

To quantitatively validate these insights we calculate eigenvector centrality, betweenness centrality, hub scores, and authority scores [47]. These measures are all run on the network described above and pictured in **Figure 2**. This is a directed network with USC titles as nodes and edges with weight (thickness) equal (and thus proportional to the thickness of the edge in the figure) to the number of sections in one title which cite sections in another title. These results are reported in the Appendix in **Supplementary Table S2–S5**.

These more formal measures of centrality largely confirm the visual impression from **Figure 2**. Title 42 has by far the highest Hub Score: twice as large as the next highest, which is Title 22 ("Foreign Relations and Intercourse"), another subject with deep connections to the rest of the Code. Title 5 has the highest Authority Score, more than three times larger than that of the second highest. The Betweenness Centrality estimates highlight the importance of two additional Titles beyond 42 and 5, which are Titles 26 ("Internal Revenue Code") and 18 ("Crimes and Criminal Procedure"). Given the importance of taxes and crime in the life of the law, substantial connections between these Titles and other parts of the law is unsurprising. The Eigenvector Centrality calculation again emphasizes the centrality of Titles 42 and 5; the other two Titles with large centrality values are 31 ("Money and Finance") and 28 ("Judiciary and Judicial Procedure"). Title 31 deals with matters such as budgeting and procurement that are cross-governmental in nature. It is also natural that the Title that deals most directly with courts connects to various other areas of the law.

4 TOPIC MODELING THE USC

Network structure is just the skeleton of the USC. The semantic content provides the meat of the law. To engage in meaningful analysis of the semantic content of statutory texts, we rely on the method of *topic modeling*, which is well suited to constructing information-rich but low-dimensional representations of large textual corpora. The following section provides a short overview of the topic modeling technique and discusses the results of a topic model applied to the USC.

Loosely speaking, a topic model is a machine learning technique that produces a description of any document in a corpus as a probability distribution (weighted sum) of a fixed (and derived) set of "topics," which should be interpreted as akin

[8]Note that this graph is based on raw count numbers of inter-title cross-references (following the procedure of counting a max of only one edge between two documents) and is not normalized by the number of documents or words in a title.

to subject matter categories [22]. Formally, a *topic* is itself a probability distribution over the vocabulary of the corpus.[9] So that in fact, a document – which in any given corpus is an *a priori* determined contiguous set of words that generally respects textual boundaries such as paragraph ends, etc. – ultimately is represented as a distribution of distributions on the vocabulary. Topics are inferred from the corpus on the basis of a set of assumptions concerning a particular kind of parametrized generative model of document construction. For a standard topic model, the parameters of interest are typically restricted to the topic-word distributions (which describe the association between topics and words) and the document-topic distributions (which describe, for each document, the probability of finding words associated with each topic). The Latent Dirichlet Allocation (LDA) model is a common prior placed on the distributions. Standard topic modeling is supervised in the sense that the number of topics is specified *a priori* rather than discovered (e.g., according to some notion of parsimonious representation). *See* [48] for a good general introduction.

We have produced a topic model representing the USC with 100 topics. A list representing each topic by its five most heavily weighted words (or word stems) is presented in the Appendix in **Supplementary Table S6**. This gives a sense of the natural subject matter category that best corresponds to a given topic. Many of the topics appear to be legally meaningful in the sense that the most heavily weighted words suggest a recognizable theme. For example, the most heavily weighted words in Topic 26 are "bank, institut, financi, feder, insur, credit." This cluster of words conforms to a legal category of banking regulation. The most heavily weighted words in Topic 99 are "educ, school, student, institut, agenc," which conform to education. Some of the topics, on the other hand, do not match substantive areas but appear to be fairly generic collections of lawmaking words: these include Topic 1 ("transfer, section, titl, codif, former") and Topic 81 ("subchapt, part, titl, section, purpose"). The prevalence of a topic in a document captures (in a useful sense) the degree to which the document is "about" each topic. More formally, the document level distribution is a latent variable that is a best fit with the observed words in the document, given a set of topic distributions. For purposes of analysis, topic prevalence is a measure of the associated semantic content of each document.

Topic prevalence can capture subtleties that would be otherwise difficult to quantitatively describe. A statutory text that could be hand-coded by a researcher might be categorized according to some set of legal subject matter categories, such as a criminal law issue or environmental law. But such issue categorizations are binary and fail to capture the mix of topics that might be present in a document. Topic prevalence, by contrast, is a set of continuous variables (i.e., representing shares for each topic) that characterize each document. *Structured topic models* (STMs) are a class of topic model that builds on this basic architecture and has been described in the peer reviewed political science literature [21].[10]

4.1 Topics and Titles

Given the description of the documents (sections) according to topic distributions, we can derive a notion of similarity between documents and with that, another kind of edge connecting document to document. In this analysis we are interested in understanding the relationship between textual similarity and cross-referencing. We first offer some intuitive illustrations. We then estimate measures of *assortativity* for the document network and consider a relational measure between content and structure inspired by the notion of *mutual information* This is based on the predictive success of algorithms using one source of data (such as topic distributions) to predict the other (such as title or cross-references).

Supplementary Table S7 (in the Appendix) provides an intuitive sense of the substantive overlap between titles and topics by showing the topic that is most closely associated with each title (as estimated via the topic share of the documents within each title).[11] Note that the numbering of the topics is arbitrary. There is a very substantial intuitive overlap between titles and their associated topics. To give just a few examples, the topic that is most associated with Title 54 ("National Park Service") has top words of "park, secretari, nation, shall, land" and the topic that is most associated with Title 38 ("Veterans' Benefits") has top words of "veteran, renumb, secretari, disabl, administer." The overlap of topics and these substantive categories nicely illustrates the power of topic models to naively discover subject matter trends within textual corpora.

Assortativity Measures

Assortative mixing in networks is "the tendency for vertices. . .to be connected to other vertices that are like (or unlike) them in some way" [49]. Assortative mixing is observed in many natural networks—for example, partisan affiliation predicts connection on social networks [50] and is akin to what is referred to in the sociological literature as *homophily* [51]. We use measures of network assortative mixing or assortativity from [49]. Assortativity is calculated over edges in a graph and is the likelihood that an edge connects two nodes with the same characteristic.

Our analysis of assortativity measures is reported in **Figure 3**. For each title, we calculate two assortativity estimates: one for title itself and the second for the topic that is most closely associated with that title (as reported in **Supplementary Table S7**). More precisely, let e_{ij} denote the fraction of document-to-document

[9]Even the term "vocabulary" is used in a somewhat non-standard fashion: a "word" is sometimes a word fragment, as per the common technique of "stemming" used in topic modeling as well as other kinds of natural language processing algorithm and the extraction of the vocabulary is more or less a standardized process.

[10]The defining feature of STMs is the ability to use metadata when constructing topics. For this paper, we do not take advantage of this feature, and so the STM we use is equivalent to the correlated topic model (CTM) described in [20], which is an extension of the LDA approach. The authors of the STM have made the model publicly available through an *R*-package at http://www.structuraltopicmodel.com/.

[11]For this analysis, we exclude several 'generic' topics that do not appear to be related to substantive legal categories. These excluded topics are: 1, 7, 16, 20, 36, 42, 43, 73, 81, and 94.

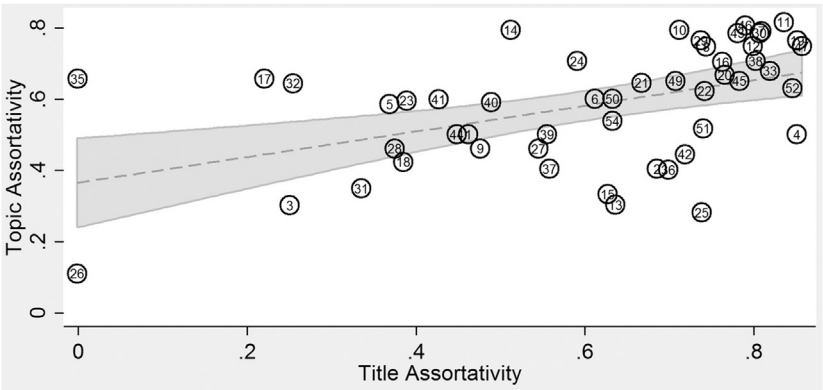

FIGURE 3 | Relationship between cross-reference assortativity at the title and topic level. Titles are matched with the topic that is most closely associated with that title as reported in **Supplementary Table S7**. Cross-references often track subject matter (as proxied by title and topic).

links (out of all links) that connect documents in titles/topics i and j (with the topic proviso above), then the *assortativity coefficient* (relative to the given labeling) is defined as

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i} \tag{1}$$

where

$$a_i = \sum_j e_{ij} \qquad b_j = \sum_i e_{ij}$$

Higher assortativity is associated with a stronger correlation between title/topic and cross-references with an estimate of 1.0 implying perfect correlation [49].

In **Figure 3** observations are titles, and are labeled as such, and the dotted line and shadow is a simple linear fit with a 95% confidence interval. As is visually apparent, there is a relationship between a title's propensity to include intra-title cross-references and for its associated topic to self-cite. This finding tends to confirm that there is an important overlap between structure and content in the USC (An OLS regression with topic assortativity as the dependent variable and title assortativity as the predictor variable shows a relatively tight relationship, with an R^2 of 0.22 and a p-value less than 0.001.)

We conduct a further analysis of the extent to which topics explain citations between titles. Our question is, for any given topic are there titles that are typically cited to from statutory sections that are closely associated with that topic?[12] We call this the *authority* measure. Slightly more formally, for a topic U and Title T we calculate $Auth(U, T)$ as the correlation between the proportion of Topic U in the source of an edge and an indicator variable for Title T as the target of that edge. This is calculated using all edges except those whose source node is in T. For each topic we took the title with the highest correlation which had a p-value of less than 0.001. The results appear to align with intuition. For example Title 21 (Food and Drugs) has the highest authority for topics 49 and 82 with top words of "control, substance, drug, chemic, test" and "product, drug, food, secretari, provid" respectively. This means that when other titles discuss these topics their citations are most likely to be to Title 21. The results are reported in **Supplementary Table S8**.

Mutual Information-like Measures

We also engage in predictive exercises to test the degree to which content and structure carry mutual information. Informally, the mutual information between two random variables attempts to measure the degree to which the observation of one random variable may assist in the prediction of the observation of a second random variable. There are formal mutual information measures, such as *Kullback-Leibler divergence*, but there are none that we are aware of that fit well with the mixed data that we are considering. Instead we use a *prediction* task to operationalize an informal understanding of mutual information: two domains of structure and content have mutual information if it is possible to use information in one domain to make predictions concerning the other. Of course, different predictive approaches (and predictive targets) may be better or worse at leveraging certain kinds of information. Nevertheless, predictive performance using actual machine-learning algorithms provides something of a sense of mutual information, as data in two completely uncorrelated spaces would not be useful for generating predictions across domains.

The first predictive task utilizes a support vector machine (SVM) algorithm trained on 40,000 (out of 41,138) randomly drawn documents with the topic proportions in the documents used to predict the title where that document was found. Testing on the held out 1, 138 documents gives an accuracy of 60%, far more than would be expected from chance.[13]

[12]For this analysis, we dropped titles that were not closely associated with any topic. These were titles 1, 3, 4, 5a, 9, 13, 14, 18a, 24, 29, 32, 35, 44, 45, and 51.

[13]The weighted average F_1 score, which accounts for both precision and recall, is 0.59. For this analysis, we relied on the scikit-learn models in Python. The model used was sklearn.svm.LinearSVC. Interestingly, we saw a very large boost in accuracy when we switched from sklearn.svm.SVC to LinearSVC, with the primary difference between the learners being the use of a linear kernel instead of a radial basis function and handling multiclass labeling as one-to-many rather than one-to-one. The relative performance of these two learners may provide insight into the underlying statutory structure.

TABLE 1 | *Search model performance at predicting cross-references.*

Method	Average performance					
	precision@10 (%)	precision@20 (%)	precision@50 (%)	recall@10 (%)	recall@20 (%)	recall@50 (%)
Proximity	7.5	3.8	1.7	2.99	3.09	3.12
Covering	16.5	11.2	5.1	6.19	8.46	12.48

Our second analysis is based on the search framework described in [16, 17]. The goal of this framework is to generate a computational model of human law search in the navigation of a multinetwork representation of a legal corpus where edges are formed through citation and semantic similarity (as instantiated via a topic model). The authors refer to this multinetwork representation as a "legal landscape." This landscape is traversed by navigating from a source document—which is an exogenously identified member of the corpus—based on two general strategies: a "proximity strategy" and a "covering strategy." The proximity strategy identifies a set of documents that are "closest" to the source document (with distance defined via a specific network-based measure described in [16] that is called PageDist – see the paper for details). The covering strategy, by contrast, attempts to "cover" the range of subjects or issues within a legal document by setting off over a related range of the landscape from the source document.

One method used to test the performance of these strategies, described in detail in [17] relies on information that is embedded in the documents. In brief, the method begins by selecting a source document and then reconstructing the landscape without that document. The source document is then stripped of citation information (leaving a "Citation Free Legal Text" or CFLT). The information in the CFLT is quite coarse-grained because semantic content is represented as topic proportions only. Based on its topic proportions the CFLT is mapped onto the legal landscape (recall that its place in the multinetwork is generated by both citation structure and semantic content, the former now removed, but the latter still intact), and the proximity or covering algorithm is deployed. The success of the model (landscape + algorithm) is tested against the actual citations that were contained in that CFLT. In [17] traditional measures of performance precision and accuracy for a given number of predicted citations are reported. (Note that the number of citations to be generated is set exogenously rather than learned through the model, under an assumption that different searchers will weigh search costs vs. information benefits differently).

[16, 17] use the opinions generated by the Supreme Court of the United States (SCOTUS) as a test corpus. Herein we extend the methodology to the USC. It is worth noting differences between SCOTUS opinions and the USC. The primary and most important difference is that the citation network for SCOTUS opinions is extremely dense, with opinions containing dozens of citations and few opinions with a very small number of citations. By comparison, the USC citation network is very sparse, with zero as the modal and median number of citations. Because citation-less documents cannot be easily incorporated into the landscape, we exclude them from this analysis. Further, because precision and recall are difficult to estimate with a very small number of citations, we limit our CFLTs to the documents that have at least five citations (for this analysis, only outgoing citations were used). It is also worth noting that the semantic content of the two corpora are very different. SCOTUS opinions are meant to persuade and generally conform to the norms of the judicial genre [5] such as stating the facts of a case in a narrative voice and offering reasons for the decision delivered. It is unclear whether the different semantic styles in the two corpora will lead to different navigation behavior on the part of law searchers.

Table 1 presents average precision and recall for different number of recommendations based on roughly 300 CFLTs. These results are roughly commensurate with the model's performance for the SCOTUS corpus (*see* [17].) For the (better performing) covering algorithm, out of the first 10 recommendations, a bit under two would be accurate matches. When the model generates 50 recommendations, a bit over 10% of the actual citations are identified.

As would be expected, there is an inverse relationship between precision and recall as the number of potential cross-references that are identified increases. An additional finding is that the covering algorithm outperforms the proximity algorithm in both the SCOTUS and USC context, indicating that it may be a more robust general approach for simulating human navigation of even very different legal corpora.

It is worth noting a further point of comparison with [17]. In that paper, comparison is made between the model's performance to human research assistants on basic research tasks, finding that although the models do not perfectly simulate natural search behavior, the degree of overlap of the model to the research assistants was not so much less than the overlap of the researchers with each other. Although we do not undertake the same analysis here, it is plausible to speculate that a similar performance would be achieved for the USC, given the model's relatively similar performance on the citation prediction task.

5 CONCLUSION AND FURTHER WORK

This preliminary analysis shows a strong association between structure and content on the USC. The former is embodied in the organizational hierarchy and section-level cross-reference link structure. The latter is quantified through the application of a structural topic modeling of the sections – a contribution in and of itself to the study of statues. There are several potential extensions that we discuss briefly in this section along with a summary of our findings.

The USC is an important corpus with profound legal effect. We find that there is a substantial degree of overlap (in the sense

of correlation) between statutory structure and content via two measures: assortativity and prediction. We find a relatively large amount of topic assortativity, as well as intuitive matches between topics and statutory titles and also strong correlations between title assortativity and the assortativity of matched topics. We also find that there is sufficient mutual information between structure and content in the sense that the topic share information of a document can be used to predict structural information (titles and cross-references) using both a trained machine learning classifier (SVM)—for titles—and via the legal landscapes approach borrowed from [16]—for cross-references.

Although these observations are interesting, it is admittedly somewhat difficult to interpret the relationships described in this paper absent a baseline for analysis. There are two possible data sources for such a baseline: domestic statutes from a comparative context, and other U.S. domestic legal orders, such as state statutory regimes. An important extension of the work reported in this paper would apply similar techniques to other statutory corpora. All of the methods described here are general and would be applicable to similarly structured statutory systems. So long as citation information can be extracted, statutory structure can be captured through the same network notions, and topic models are language-agnostic so long as the underlying texts are machine readable. What we describe in this paper then amounts to an off-the-shelf methodology that can, in principle be applied to any statutory regime. Comparative work along these lines is likely to be particularly fruitful because it provides a means of examining the relative strength of the relationship between structure and content in different statutory regimes. We cannot know from this initial analysis whether the United States is an outlier, or if similarly situated legal orders tend to have similar levels of content/structure interrelatedness. In another direction, comparative work of state-level statutory networks would also be interesting.

A second extension would delve deeper into the USC itself, perhaps by linking this corpus to other legal texts, such as the Code of Federal Regulations. Especially with a larger total corpus, it would be possible to make intra-USC comparison between, for example, different titles or different substantive areas (as estimated either naively via a topic model or through expert labeling). It would be worth investigating whether some areas of the law are more "self-contained" than others, and whether any measure of self-containment correlated with other characteristics, such as the size of the industry that was regulated or partisan dynamics (either temporally at the national level or geographically at the state level).

A final set of extensions are more practical in nature. Statutory texts are notoriously difficult to navigate, in part because they lack the kind of identifying information that is contained in judicial opinions. Research into the relationship between statutory structure and content (and the relationships between statutes and other legal documents) could be used as the foundation for new search/navigation tools. Such tools could be used by practitioners to lower the transaction costs associated with identifying relevant statutory texts. Given the substantial private expenditures on law search, any technique that lowered those costs by even a small percentage would create a substantial amount of economic value.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://uscode.house.gov/.

AUTHOR CONTRIBUTIONS

FD and KC did much of the data extraction and computation. All authors analyzed the results. All authors contributed to the writing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphy.2020.625241/full#supplementary-material.

REFERENCES

1. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. Multilayer networks. *J Complex Networks* (2014) 2(3):203–71. doi:10.1093/comnet/cnu016
2. Bommarito MJ, Katz DM. A mathematical approach to the study of the United States Code. *Physica A: Stat Mech Appl* (2010) 389(19):4195–200. doi:10.1016/j.physa.2010.05.057
3. Corominas-Murtra B, Goñi J, Solé RV, Rodríguez-Caso C. On the origins of hierarchy in complex networks. *Proc Natl Acad Sci* (2013) 110(33):13316–21. doi:10.1073/pnas.1300832110
4. Mengistu H, Huizinga J, Mouret JB, Clune J. The evolutionary origins of hierarchy. *PLoS Comput Biol* (2016) 12:e1004829. doi:10.1371/journal.pcbi.1004829
5. Livermore MA, Riddell AB, Rockmore DN. The Supreme court and the judicial genre. *Ariz L Rev* (2017) 59(4):837–901.
6. Rockmore DN, Fang C, Foti NJ, Ginsburg T, Krakauer DC. The cultural evolution of national constitutions. *J Assoc Inf Sci Technol* (2017) 69(3):483–94. doi:10.1002/asi.23971
7. Katz DM, Bommarito MJ. Measuring the complexity of the law: the United States Code. *Artif Intell L* (2014) 22(4):337–74. doi:10.1007/s10506-014-9160-8
8. Sunstein CR. *Simpler: the future of government*. New York, NY: Simon & Schuster (2013).
9. Al-Ubaydli O, McLaughlin PA. RegData: a numerical database on industry-specific regulations for all United States industries and federal regulations, 1997–2012. *Regul Governance* (2017) 11(1):109–23. doi:10.1111/rego.12107
10. Carey, MP. Counting regulations: an overview of rulemaking, types of federal regulations, and pages in the federal register (2016). Congressional Research Service. Technical report.
11. Lumsdaine RL, Rockmore DN, Foti NJ, Leibon G, Farmer JD. The intrafirm complexity of systemically important financial institutions. *J Financial Stab* (2020) 52:100804. doi:10.1016/j.jfs.2020.100804

12. Fowler JH, Johnson TR, Spriggs JF, Jeon S, Wahlbeck PJ. Network analysis and the law: measuring the legal importance of precedents at the U.S. Supreme Court. *Polit Anal* (2007) 15(3):324–46. doi:10.1093/pan/mpm011
13. Rice D. The impact of Supreme court activity on the judicial agenda. *L Soc'y Rev* (2014) 48:63–90. doi:10.1111/lasr.12056
14. Carlson K, Livermore MA, Rockmore DN. A quantitative analysis of writing style on the U.S. Supreme court. *Wash U L Rev* (2016) 93:1461.
15. Carlson K, Livermore MA, Rockmore DN. The problem of data bias in the pool of published U.S. Appellate court opinions. *J Empirical Leg Stud* (2020) 17(2): 224–61. doi:10.1111/jels.12253
16. Leibon G, Livermore M, Harder R, Riddell A, Rockmore D. Bending the law: geometric tools for quantifying influence in the multinetwork of legal opinions. *Artif Intell L* (2018) 26(2):145–67. doi:10.1007/s10506-018-9224-2
17. Livermore MA, Rockmore DN, Dadgosari F, Guim M, Beling PA. Modeling law search as prediction. *Artifi Intell Law* (2020) 29:3–34. doi:10.1007/s10506-020-09261-5
18. Postema GJ. *Bentham and the common law tradition*. Oxford, United Kingdom: Oxford University Press (1989).
19. Badawi AB, Dari-Mattiacci G. Reference networks and civil codes In: MA Livermore DN Rockmore, editors *Law as data: computation, text, and the future of legal analysis*. Chap. 12, San Fante, NM: SFI Press (2019). p. 339–65.
20. Blei DM, Lafferty JD. A correlated topic model of Science. *Ann Appl Stat* (2007) 1(1):17–35. doi:10.1214/07-aoas114
21. Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, et al. Structural topic models for open-ended survey responses. *Am J Polit Sci* (2014) 58(4):1064–82. doi:10.1111/ajps.12103
22. Quinn KM, Monroe BL, Colaresi M, Crespin MH, Radev DR. How to analyze political attention with minimal assumptions and costs. *Am J Polit Sci* (2010) 54(1):209–28. doi:10.1111/j.1540-5907.2009.00427.x
23. Riddell AB. How to read 22,198 journal articles: studying the history of German studies with topic models In: M Erlin L Tatlock, editors *Distant readings: topologies of German culture in the long nineteenth century*. Rochester, NY: Camden House (2014). p. 91–114.
24. Jockers ML, Mimno D. Significant themes in 19th-century literature. *Poetics* (2013) 41(6):750–69. doi:10.1016/j.poetic.2013.08.005
25. Roe G, Gladstone C, Morrissey R. Discourses and disciplines in the enlightenment: topic modeling the French encyclopédie. *Front Digit Humanit* (2016) 2:8. doi:10.3389/fdigh.2015.00008
26. Schöch C. Topic modeling genre: an exploration of French classical and enlightenment drama. *DHQ: Digital Humanities Q* (2017) 11(2):266–85. doi:10.5281/zenodo.166356
27. Macey J, Mitts J. Finding order in the morass: the three real justifications for piercing the corporate veil. *Cornell L Rev* (2014) 100(1):99–155.
28. Law DS. Constitutional archetypes. *Tex L Rev* (2016) 95(2):153–243.
29. Lauderdale BE, Clark TS. Scaling politically meaningful dimensions using texts and votes. *Am J Polit Sci* (2014) 58(3):754–71. doi:10.1111/ajps.12085
30. Chalfin A, Haviland AM, Raphael S. What do panel studies tell us about a deterrent effect of capital punishment? A critique of the literature. *J Quant Criminol* (2013) 29(1):5–43. doi:10.1007/s10940-012-9168-8
31. Graham ER, Shipan CR, Volden C. The diffusion of policy diffusion research in political science. *Br J. Polit. Sci.* (2013) 43(03):673–701. doi:10.1017/s0007123412000415
32. La Porta R, Lopez-de SIlanes F, Shleifer A, Vishny RW. Law and finance. *J Polit Economy* (1998) 106(6):11131–55. doi:10.1086/250042
33. Porta RL, Lopez-de-Silanes F, Shleifer A. The economic consequences of legal origins. *J Econ Lit* (2008) 46(2):285–332. doi:10.1257/jel.46.2.285
34. Michaels R. Comparative law by numbers? Legal origins thesis, doing business reports, and the silence of traditional comparative law. *Am J Comp Law* (2009) 57(4):765–95. doi:10.5131/ajcl.2008.0022
35. Klerman DM, Mahoney PG, Spamann H, Weinstein MI. Legal origin or colonial history?. *J Leg Anal* (2011) 3(2):379–409. doi:10.1093/jla/lar002
36. Spamann H. Empirical comparative law. *Annu Rev L Soc. Sci.* (2015) 11: 131–53. doi:10.1146/annurev-lawsocsci-110413-030807
37. Ruhl JB, Katz DM. Measuring, monitoring, and managing legal complexity. *Iowa L Rev* (2015) 101(1):191–244.
38. Schuck PH. Legal complexity: some causes, consequences, and cures. *Duke L J* (1992) 42(1):1–52. doi:10.2307/1372753
39. Tainter JA. *The collapse of complex societies*. Cambridge, United Kingdom: Cambridge University Press (1988).
40. Fonseca R, Lopez-Garcia P, Pissarides CA. Entrepreneurship, start-up costs and employment. *Eur Econ Rev* (2001) 45(4-6):692–705. doi:10.1016/s0014-2921(01)00131-3
41. Nicoletti G, Scarpetta S. Regulation, productivity and growth: OECD evidence. *Econ Policy* (2003) 18(36):9–72. doi:10.1111/1468-0327.00102
42. Ciccone A, Papaioannou E. Red tape and delayed entry. *J Eur Econ Assoc* (2007) 5(2-3):444–58. doi:10.1162/jeea.2007.5.2-3.444
43. Braunerhjelm P, Eklund JE. Taxes, tax administrative burdens and new firm formation. *Kyklos* (2014) 67(1):1–11. doi:10.1111/kykl.12040
44. Ellig J, McLaughlin PA. The regulatory determinants of railroad safety. *Rev Ind Organ* (2016) 49(2):371–98. doi:10.1007/s11151-016-9525-0
45. Chambers D, Collins CA, Krause A. How do federal regulations affect consumer prices? An analysis of the regressive effects of regulation. *Public Choice* (2017) 180(1):57–90. doi:10.1007/s11127-017-0479-z
46. U.S. Chamber of Commerce. The regulatory impact on small business: complex. cumbersome. costly (2017). U.S. Chamber of Commerce. Technical report.
47. Kleinberg JM. Authoritative sources in a hyperlinked environment. *J Acm* (1999) 46(5):604–32. doi:10.1145/324133.324140
48. Blei DM. Probabilistic topic models. *Commun ACM* (2012) 55(4):77–84. doi:10.1145/2133806.2133826
49. Newman MEJ. Mixing patterns in networks. *Phys Rev E* (2003) 67(2). doi:10.1103/physreve.67.026126
50. Colleoni E, Rozza A, Arvidsson A. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in twitter using big data. *J Commun* (2014) 64(2):317–32. doi:10.1111/jcom.12084
51. McPherson M, Smith-Lovin L, Cook JM. Birds of a feather: homophily in social networks. *Annu Rev Sociol* (2001) 27(1):415–44. doi:10.1146/annurev.soc.27.1.415
52. Whisner M. The United States code, prima facie evidence, and positive law. *L Libr Journal* (2009) 101(4):545–56.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.