Network Measures of the United States Code

Alexander Lyte
The MITRE Corporation
7515 Colshire Dr.
McLean, VA
alyte@mitre.org

David Slater
The MITRE Corporation
7515 Colshire Dr.
McLean, VA
dslater@mitre.org

Shaun Michel
The MITRE Corporation
7515 Colshire Dr.
McLean, VA
smichel@mitre.org

ABSTRACT

The US Code represents the codification of the laws of the United States. While it is a well-organized and curated corpus of documents, the legal text remains nearly impenetrable for non-lawyers. In this paper, we treat the US Code as a citation network and explore its complexity using traditional network metrics. We find interesting topical patterns emerge from the citation structure, and begin to interpret network metrics in the context of the legal corpus. This approach has potential for determining policy dependency and robustness, as well as modeling of future policies.

Categories and Subject Descriptors

D.3.3 [**Programming Languages**]: Python, Neo4j, Cypher, Javascript; graph theory

General Terms

Graph Theory; Legal Analysis

Keywords

Policy networks, rulesets, United States Code

1. INTRODUCTION

The US Code (USC) is a large, complex, interconnected corpus of laws that regulate much of American life. With laws regulating the Armed Forces, Conservation, Banking, and much more, it not only is an interesting dataset from a semantic perspective, but also has the potential to reveal interesting aspects about the US legal regulatory space.

In this paper, we treat the USC as a citation network, and analyze it using traditional network approaches. We explore some key phenomenon, including the density of connections, the interrelations among titles, and the emergence of community structures within the graph.

In section 2, we review previous work on parsing laws, analyzing their text content, and building citation networks of interrelated legal documents. Section 3 gives an overview of the USC, its generation process, organizational structure, publicly available forms, and an overview of how we construct our citation network. Section 4 walks though baseline metrics on the graph, including number of nodes, edges, degree, betweenness, and centrality by title. Section 5 explores the interdependencies among the titles, and section 6 describes the results of community detection testing on the graph. Section 7 concludes with a discussion of future directions.

2. RELATED WORK

Building citations of legal text is not new. Koniaris outlines some of the research done on various legal corpora and takes a computational approach to parsing the legal text, defining some standard reference types such as ”amended by”, ”legal basis”, and ”instruments cited”[4]. The article provides a framework for integrating various document corpora, such as treaties, legislation, and jurisprudence, and explores some subgraphs of European Union legislation. Several network metrics are established, including degree distribution, node-edge ratios, and resiliency.

Katz & Bommarito explore the USC from the perspective of knowledge acquisition, ”a field at the intersection of psychology and computer science”[3]. In doing so, they provide metrics on the structure, linguistic content, and interdependence of the USC Titles.

In ”Towards Automated International Law Compliance Monitoring”, Morgenstern explores the feasibility of parsing text at the sentence level for use in a rule template framework[5]. Her work develops an architecture for bulk processing of legal text, but notes the challenges in parsing bulleted text in the Irrealis mood. In parsing the text, Morgenstern notes that there are different classes of citations, breaking them down in several ways. First, definitions are treated as their own type of citation. These are used to build an ontology of terms. Second, regulatory citations are classified as either cross-document, intra-document, or branch, depending on where the cited document lives. Third, exemptions are classified in a way that allows for formalization. Lastly, regulation types are identified as ”obligations, permissions, prohibitions, penalties, and reparations.”. These classifications help to codify the law into a set of business rules and processes.

This recent research helps to clarify the network structure and content, and provides a framework for mapping the functions of the law. With these insights and guiding metrics, we further explore how the function of the law can be understood by its network structure

3. THE UNITED STATES CODE

The USC represents the compiled federal statutory law of the United States. It is published every six years by the Of-

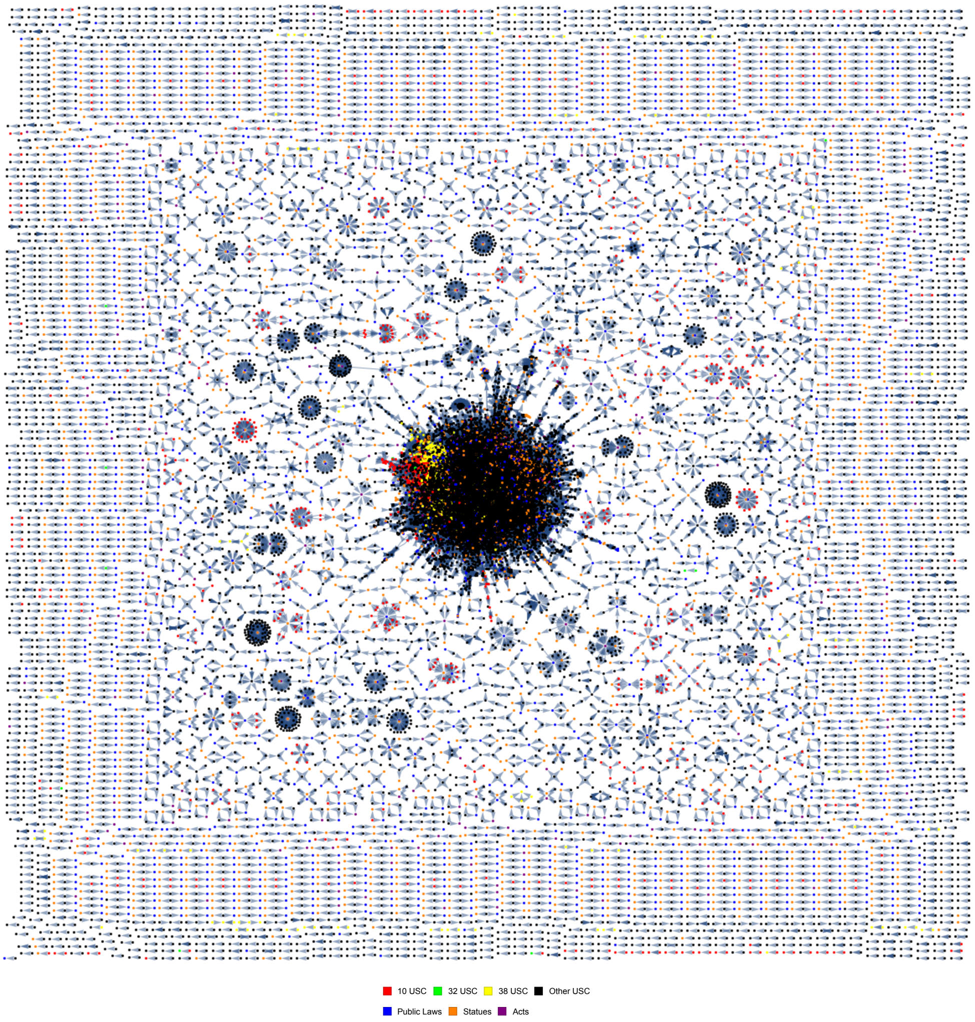

Figure 1: Network of References of US Code at Section Level

fice of the Law Revision Counsel of the House of Representatives, with cumulative supplements published annually. The laws in the USC span all aspects of government, from the Executive, Legislative, and Judicial branches, to the rights of citizens, the duties of agencies, and much more.

3.1 Derivation of Laws

The laws in the USC are a product of the legislative branch of the US government. Bills are voted on and approved by congress and then sent to the President for signature. Once signed, the bill is delivered to the Office of the Federal Registrar for authorization. Once authorized, copies are distributed as 'slip laws' by the US Government Publishing Office (GPO). The GPO Archivist assembles volumes of laws annually and publishes them as US Statutes at Large. The Office of Law Revision Counsel of the US House of Representatives restates the texts of statutes to mitigate ambiguity and obsolescence while accurately reflecting the original effects of each document[6].

3.2 Hierarchy/layout breakdown

The USC has 52 active Titles organized generally by topic. Each Title is subdivided into a series of topics related to the Title, though not necessarily in any ontological hierarchy. There are 14 structural levels in the document, each of which can be invoked by reference. These levels are hierarchical, ranging from the Title-as-object as the most macro-level container of data to various sub-items (e.g. chapters, paragraphs, clauses, bulleted items, etc.) which constitute the document's content. Not all titles use all series of subdivisions, but ”sections are of particular importance because they are both the first level at which substantive text appears and the first level at which the hierarchy can terminate”[3]. The USC can be modeled both as a hierarchical network (vertically) from title to section to subsection, and as a citation network (horizontally) connecting sections that reference each other. In our analysis, we model the USC by treating sections as nodes in our graph.

3.3 Digital Versions of the USC

The Office of the Law Revision Counsel develops an Extensible Markup Language (XML) version of the USC, which

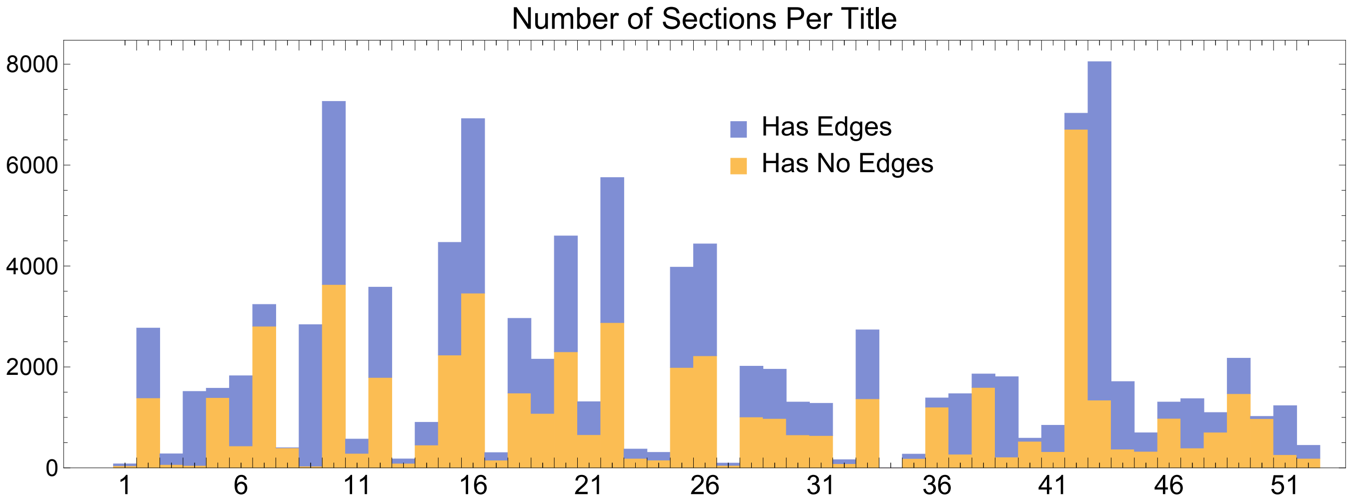

Figure 2: Network of References of US Code at Section Level

the GPO makes available online. The XML format is designed to be consistent with the Akoma Ntoso project, which is an international standard for legal text markup. The XML includes explicit reference citations where possible, which we extract and use to create the citation network.

3.4 Citation Network Construction

We use XML data downloaded from the Office of the Law Revision Counsel site to build a citation network. Our data is from Release Point 113-271, released December 18, 2014. It includes Titles 1-52, excluding Title 34, which remains absent from the list of USC Titles at the time of this paper's authorship. Title 53 was reserved for future use and Title 54 was excluded because it was included with the subsequent Release Point.

For each document, a Python script parses the XML tree and extracts section identifiers, textual references, and sectional texts from the XML. This process allows us to encode each node according to a ready-made schema seen in the XML in which a Uniform Resource Locator (URL) in the form of */Text/Title/Section/* is used for every section. References, encoded in *<ref>* tags, are pulled out and used to build an edge list, with the analyzed section as the source and the referenced section the target of the citation relationship. As a result, we create a single dataset containing an adjacency edgelist and relevant metadata to construct a directed network for the entire USC. Referenced sections (targeted nodes) are not limited to USC, but can include sections found in Code of Federal Regulations documents and Public Laws; these other documents have not been parsed, so they do not produce any source nodes in our dataset.

4. GENERAL NETWORK METRICS

Our citation network initially includes 67,286 nodes from the US Code. In the following graph analyses, we exclude from our graph sections of the USC with no edges either in or out (nodes with degree count equal to 0), resulting in 33,239 USC non-isolated nodes being considered. This yields 92,166 in-text references between documents, intentionally ignoring references to non-USC documents and information in post-section notes (e.g. general references to repeals, amendments, or enactment dates).

Figure 1 shows the full graph of the USC as we define it. While there are many familiar graph structures, such as cycles, stars, and pairs, the most prominent feature is a gigantic weakly connected componentâĂŤmeaning that all nodes in the defined component have an observed path to all other nodes in the componentâĂŤlocated in the center. This component contains many interwoven parts of the USC, but can also be broken into communities, as we explore in Section 6 of this paper.

Many of the disjointed components contain a single center node, sometimes in a star-shape, other times connecting multiple subgraphs. In traditional graph theory, the structure of these networks can provide insight into the actual system. In the rest of this section, we explore some traditional graph metrics and explore how they relate to the laws they describe.

4.1 Number of Nodes and Edges

We begin with a count of section nodes in our graph, which varies dramatically across titles. Figure 2 illustrates the total number of nodes in each title, and shows the ratio between nodes with no edges or singletons (0 degree nodes) and nodes with edges or connected nodes. Note, for example, that title 43 has a high proportion of connected nodes, while title 42 primarily contains isolates, although both contain a very large number of sections.

The number of edges in this citation network speaks to the number of times that sections explicitly reference other sections. Within the context of the USC, citations can be used to provide a definition, establish authority, note exceptions, and even to repeal laws. By importing text from other sections, the total amount of information within a section can be vastly increased in an efficiently scalable way.

In Figure 3, we break down the number of references across all sections within each title, indicating the log ratio between total in-degree and out-degree (e.g. number of times a section is referenced and number of references a section makes, respectively). While the total number of citations varies across titles, the in- and out-degree counts

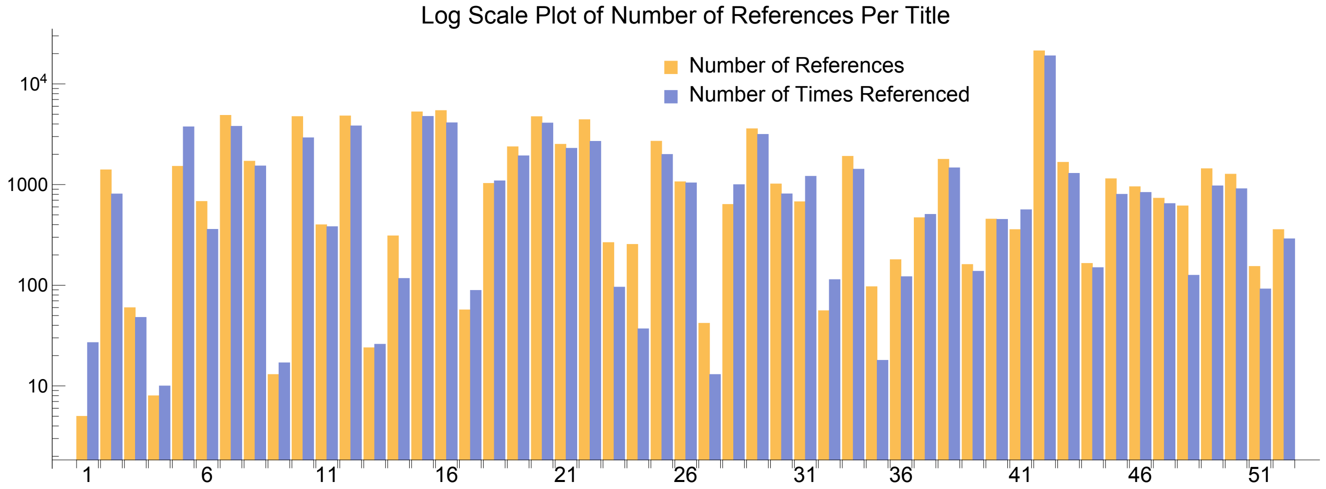

Figure 3: Network of References of US Code at Section Level

generally appear to approximate each other's magnitude for any title. Additionally, we see several titles with high reference counts and low section counts, indicating that some sections have very high reference value.

4.2 In/Out Degree

A node's degree is the number of edges going into or out of it. As shown in Figure 4 (top), the median degree of most titles is close to 1, but almost all titles have large outliers. This non-normal distribution of node degree is characteristic of graphs with powerlaw distributions. Other examples of systems with similar distributions include internet web traffic and the population of cities, among many others[1]. Figure 4 displays box-and-whisker plots and highlights a small number of very high-degree nodes in each title (note the log scale).

Table 1: Top five most cited nodes

Section	In-degree	Section Title
8 USC 1011	447	Definitions
42 USC 1395x	385	Definitions
5 USC 552	376	Public information; agency rules, opinions, orders, records, and proceedings
5 USC 553	274	Rule Making
12 USC 1813	261	Definitions

Nodes with a high in-degree are cited by many sections, while nodes with high out-degree cite many sections. The top 5 in-degree (most cited) nodes are shown in Table 1. "Definitions" sections often have high in-degrees. Many sections pull in definitions to establish consistent meanings across contexts. Definitions can be entities (such as "employee", "vehicle", or "agency") and concepts (such as "retirement age") and are used within and across topical sections. Nodes with high out-degree cite many other sections of the USC. The top 5 out-degree nodes (most citations) are shown in Table 2. The sections in Table 2 all deal with complex topics that involve many specific cases, conditions, and exceptions. They cite many other parts of the law for precision and clarification. One interpretation of this may be that high in-degree nodes provide useful information to many sections, while high out-degree nodes receive information from many sections.

Table 2: Top five most cited nodes

Section	Out-degree	Section Title
42 USC 1396a	251	State plans for medical assistance
42 USC 13951	203	Supplementary Medical Insurance Benefits for Aged and Disabled - Amounts
8 USC 1184	142	Admission of Non-Immigrants
8 USC 1182	138	Inadmissible Aliens
42 USC 402	138	Old-age and survivors insurance benefit payments

4.3 Betweenness Centrality

In network terms, a path is any collection of nodes that may be traversed to connect one particular node to another. The shortest path is one which requires the fewest intermediary nodes to complete the path. There may be several such paths of equal length and each one is important in the betweenness centrality metric. The unnormalized betweenness centrality metric of a node k measures the number of shortest paths that rely upon k to properly connect any node i with any other node j. In other words, the metric counts how frequently k is on the shortest paths between all pairs of nodes in the network.

Because our network is directed, a section with high betweenness centrality suggests that it is not only important for establishing context for a variety of other sections but may also draw upon a variety of other sections to establish its own context. The responsible section node is the source of a bottleneck effect, but, without it, there would be an ab-

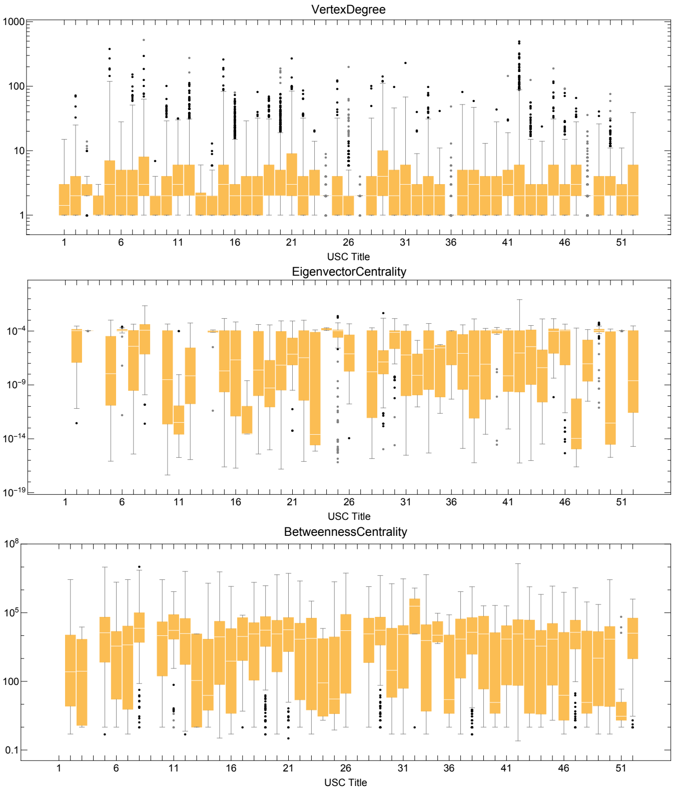

Figure 4: Box-Whisker Charts for Vertex Degree, Eigenvector Centrality, and Betweeness Centrality for each Title of the US Code

sence of connection between the different groups of nodes or else many of the nodes that currently reference the one with a high betweenness would instead have to reference many of those that it references to create the same effect of context.

Commonly, betweenness metrics in citation networks can reveal the nature of interdisciplinary interactions because, for example, specialists may intrinsically link together while generalists operate between specialist groups. For our purposes, this chaining of references is indicative of the interdependencies between titles in a hierarchical manner: some titlesâĂŤindeed, some specific sectionsâĂŤmay be more or less critical to establishing context for other titles.

It is possible that, given the existence of betweenness, a section must produce a legal effect that is then built upon by those which refer to it, or else it wouldn't be referred to at all and each reference link would instead go to a more appropriate source.

Figure 4 (middle) depicts a box-and-whisker plot that shows the distribution of the collective range of betweenness centrality values for all sections within each of the fifty-one active USC titles we examined. As to be expected, larger USC documents such as titles 5, 26, and 42âĂŤwhich also contain more sectionsâĂŤtend to have a larger distributional range of betweenness centralities, as seen in Table 3.

4.4 Eigenvector Centrality

An eigenvector centrality metric represents a node's connectivity to other highly connected nodes. A higher value suggests that it is somehow important to nodes which we might intuitively expect to be important themselves. In a directed graph, such as ours, the in-degree (edges pointing towards a node, rather than away from) is critical to determining the eigenvector value. For our purposes, a section node i which is referenced by many other sections (a high in-degree value) will make another node k appear more important simply by i referencing k. This is different from betweenness centrality because k does not need to reference anything to maintain a higher eigenvector centrality status; this is a potential metric of a node's overall relative importance by being highly depended upon. The distribution of eigenvector centrality for each title is shown in Figure 4(bottom). These raw values can be compared across nodes to identify the more central among them. As shown in Table 4, Title 8 contains the most central node by this metric (8 USC 1101).

Table 3: Top five sections by betweeness centrality

Section	Betweeness	Section Title
42 USC 402	1.40E+07	Old-age and survivors insurance benefit payments
8 USC 1101	1.05E+07	Definitions (Immigration & Nationality)
5 USC 552a	9.92E+06	Records Maintained on Individuals
8 USC 1182	7.23E+06	Inadmissible Aliens
12 USC 1813	6.39E+06	Definitions (FDIC)

Table 4: Top five sections by betweeness centrality

Section	Eigenvector	Section Title
8 USC 1101	532	Definitions (Immigration & Nationality)
42 USC 1396a	505	State plans for medical assistance
42 USC 1395x	464	Definitions
5 USC 552	381	Public information; agency rules, opinions, orders, records, and proceedings
42 USC 402	329	Old-age and survivors insurance benefit payments

5. TITLE INTERDEPENDENCY

Some titles, such as Titles 5 and 42, are not only highly cited, but act as important pathways between parts of the law. In this model of the law as an interconnected network, connections between titles are of particular interest. One would expect that similar topics are well-connected, and may draw from similar sources. In network theory, an adjacency matrix is used to show which nodes are connected to which other nodes. Since we are dealing with nodes at the title level, we can aggregate the percentage of connections from one title to another across all sections. In Figure 7, we

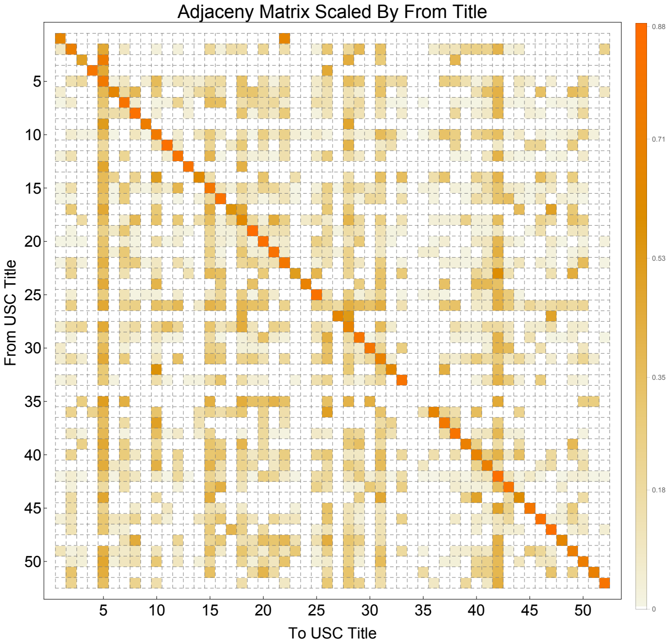

Figure 5: Adjacency Matrix of Title References

display a heat map of the aggregated adjacency matrix of the USC.

Within this heat map Titles 5 and 42 span their entire columns, showing that almost every other title makes references to 5 and 42. Alternately, Titles 13 and 27 reference themselves almost exclusively. The USC is organized topically, but topics often have overlapping elements. Title 5 deals with ”Government Organizations and Employees”, which are concepts that are used in describing ”The President” (Title 3), the ”Armed Forces” (Title 10), and the ”Postal Service” (Title 39). Title 42 deals with ”The Public Health and Welfare”, a concept that has implications from ”Highways” (Title 23) to ”Voting and Elections” (Title 52).

With this high-level information, one could infer that Title 32 is composed of content related to Titles 10, 37, 42, and 52. This implies that the ”National Guard” topic is a composite of text dealing with ”The Armed Forces”, ”Pay and Allowances of the Uniformed Services”, ”The Public Health and Welfare”, and of course ”The National Guard”, since it cites other parts of Title 32. It would take further analyses to deduce what specific concepts and entities are drawn from which sources, and which ones are unique to the specific topic. However, the notion that a given title is a composite of a select set of other titles leads to the question: are there groups of titles which cite each other strongly, and only cite other topics weakly or not at all? This notion of higher-level ”groupings of topics” is akin to the network technique of ”community detection”, which we explore in the next section.

6. COMMUNITY METRICS

The final network analysis technique we explore is community detection. Communities are defined as groups of densely interconnected nodes that are only sparsely connected with the rest of the network. There are several well-defined algorithms for community detection, each with advantages and disadvantages[2]. Given the size of the USC graph, and the computational requirements for each technique, only the Modularity-based community detection algorithm was able to converge on a solution in a reasonable amount of time. For this reason, we will interpret our findings solely though this lens.

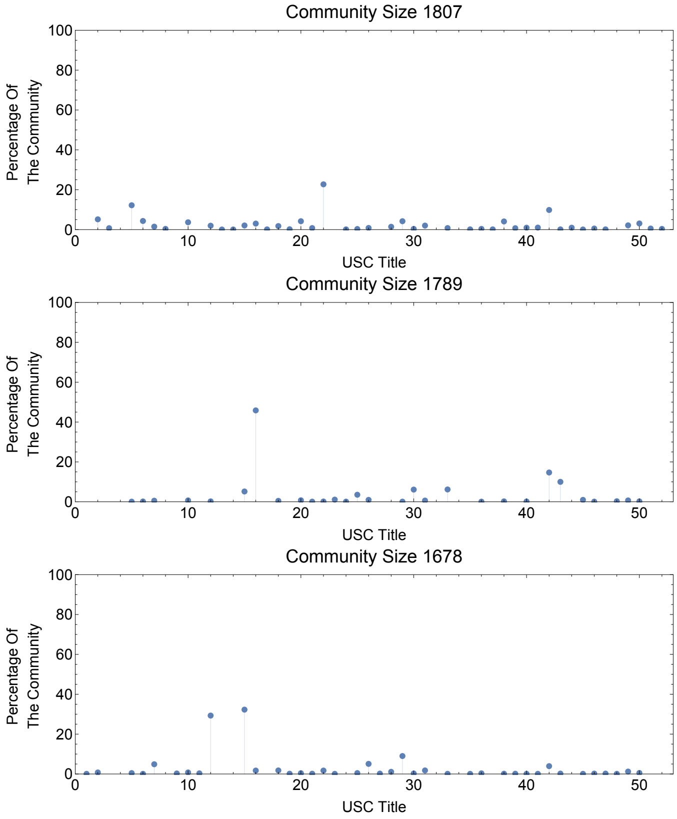

Figure 6: Top three largest communities in the United States Code

Modularity-based community detection is an approach to algorithmically finding ”communities” in a graph. The idea is to take a given group of nodes and test whether the concentration of edges within the module is greater than what would be expected in a random distribution of edges between all nodes regardless of the modules. This approach segments a graph into dynamically generated non-overlapping ”communities”, which are comprised of groups of nodes of various sections.

We are interested in applying the modularity algorithm on the largest connected component of the graph. Given that the USC is organized topically, one might expect the graph to segment cleanly into titles. Alternatively, given the commonalities between titles such as ”The Armed Forces” and ”The National Guard”, higher-level topics may group together as well. In our analysis, we see a little of both effects.

In figure 8, we see the distribution of titles in the largest 3 communities. The largest community detected is composed primarily of Title 22 (Foreign Relations and Intercourse), but also includes a large amount of Titles 5 (Government

Organization and Employees) and 42 (The Public Health and Welfare). The second largest component is chiefly composed of Title 16, (Conservation), with some of 42 and 43 (Public Lands). In the third, Titles 12 (Banks and Banking) and 15 (Commerce and Trade). While not all of the communities are so clear-cut, these first three do illustrate an aggregation of higher-level concepts.

7. DISCUSSION AND CONCLUSIONS

In analyzing the USC as a citation network, we find several interesting phenomena. A raw count of sections by title reveals vast differences in size across titles, and the count of edges shows evidence of strong interrelationships among them. The degree distribution of section nodes is skewed heavily for all titles, with most nodes having one or fewer connections, and a few with hundreds. This data-focused approach quickly teases out some of the most cited nodes, providing a quick heuristic for relative importance. Further analysis can give context for why some nodes have such tremendous import, and may provide a basis for a more semantic interpretation of the graph.

Betweenness and eigenvector centrality metrics have been derived for each title, and we have begun to interpret their usefulness in this context.These metrics speak to the major pathways of references from one section to another. With further work in understanding the types of references, one may find chains of meaning in one form or another. Further research must be done to determine the validity and utility of such metrics.

By exploring the heat map of the adjacency matrix, we saw that almost every title cites Titles 5 and 42, while Titles such as 13 and 27 cite almost no other parts of the law. This led to the hypothesis that groups of laws regulating similar topics are more closely connected because they deal with similar concepts and cite similar codes. We began to see this by using a community detection technique to determine statistically significant communities within the US Code's biggest component. By looking over several detected communities, we found that in most cases, the community detection algorithm subdivided the graph almost completely by title, producing components dominated by single titles. However, the algorithm also grouped like titles, showing how two related titles share a similar legal basis. These relationships may be useful for providing frameworks for new laws, by showing where certain topical constructs are derived and constructed.

While still very experimental, the network representation of the USC has obvious potential in legal dependency analysis and may have applications in policy modeling. Overall, this has is a rich data source to test graph metrics on and may be of interest to ontologists, linguists, policy analysts, and others.

8. REFERENCES

[1] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[2] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3–5):75 – 174, 2010.

[3] D. M. Katz and I. Bommarito, M.J. Measuring the complexity of the law: the united states code. *Artificial Intelligence and Law*, 22(4):337–374, 2014.

[4] M. Koniaris, I. Anagnostopoulos, and Y. Vassiliou. Network analysis in the legal domain: A complex model for european union legal sources. *CoRR*, abs/1501.05237, 2015.

[5] L. Morgenstern. *Toward Automated International Law Compliance Monitoring (TAILCM)*, 2014.

[6] O. of the Law Revision Counsel. Process of positive law codification.