
This study investigates the inherent structure of the network of citations to judicial precedent among North Carolina appellate judicial decisions, and seeks to determine whether there exist natural groupings or communities within that network that correlate with particular legal topics. In the study, a network is constructed from the texts of North Carolina Supreme Court and Court of Appeals decisions issued between 1998 and 2011 and analyzed by means of an algorithm designed to identify community structure within complex social networks.

The results of the analysis suggest that the inherent structure of a network of citations to judicial precedent might be used to group topically related judicial decisions together. However, the representation of a judicial decision as a single node in such a network is potentially inconsistent with the fact that a single judicial decision may address multiple disparate legal issues, each with its own independent citations to judicial precedent.

Headings:

Citation analysis

Citation analysis -- Case studies

Citation analysis -- Evaluation

Law

Legal Research
IDENTIFYING COMMUNITY STRUCTURE INHERENT IN THE CITATION NETWORK OF JUDICIAL PRECEDENT

by
James W. Vaughan

A Master’s paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Information Science.

Chapel Hill, North Carolina
November 2011

Approved by

______________________________
Robert Losee
Table of Contents:

**Judicial Precedent and Stare Decisis** ................................................................. 4

  Topical classification of judicial decisions. ......................................................... 6

**Network Analysis of Judicial Precedent** ............................................................ 7

  The validity of judicial precedent. ............................................................................. 7

  Citation analysis ......................................................................................................... 9

  Judicial precedent as a network. ............................................................................. 13

  Creation of a network of judicial precedent. ......................................................... 15

  Network analysis of citations to judicial precedent. ............................................. 19

**Community Structure in Networks of Judicial Precedent** .................................. 21

  Identifying community structure in networks. ....................................................... 21

  Community structure in citation networks. ............................................................. 26

**Research Question and Hypotheses** .................................................................. 29

**Method** .................................................................................................................. 30

**Collection of Data** ............................................................................................... 30

  North Carolina appellate jurisprudence. ................................................................. 30

  Sources of data. ......................................................................................................... 31

**Parsing of Data** ..................................................................................................... 36

  Goal of parsing. ....................................................................................................... 36

  Parsing efforts. ......................................................................................................... 36

**Detection of Communities** .................................................................................. 43

  Perl implementation. ............................................................................................. 43
R/igraph library. ................................................................. 45

Results .................................................................................. 46

Pilot Network........................................................................... 46

Communities generated. ......................................................... 46

Full Network........................................................................... 48

Communities generated. ......................................................... 49

Community content. ............................................................... 49

Discussion ............................................................................. 51

Community Detection Algorithms ......................................... 51

Communities within a Network of Citations to Judicial Precedent .................................................. 52

Conclusion ............................................................................ 53

Potential Impact of Research ................................................... 54

Legal research. ....................................................................... 54

Legal theory. .......................................................................... 54

Improvements to Current Research .......................................... 55

Judicial decision data. ............................................................. 55

Judicial decision parser. .......................................................... 56

Output analysis........................................................................ 58

Further Research ................................................................... 60

Weighted edges. ................................................................... 60

Nodal statutes. ....................................................................... 62

Divided nodes. ....................................................................... 62

Overlapping communities. ...................................................... 64

Expanded network. ............................................................... 64

Legal research. ....................................................................... 54

Legal theory. .......................................................................... 54

Improvements to Current Research .......................................... 55

Judicial decision data. ............................................................. 55

Judicial decision parser. .......................................................... 56

Output analysis........................................................................ 58

Further Research ................................................................... 60

Weighted edges. ................................................................... 60

Nodal statutes. ....................................................................... 62

Divided nodes. ....................................................................... 62

Overlapping communities. ...................................................... 64

Expanded network. ............................................................... 64
Identifying Community Structure Inherent in the
Citation Network of Judicial Precedent

In studying the social relationships among members of a large group of people, the role of a particular member in the context of the broader group dynamic can be studied by looking at the social connections between that member and the other members of the group. By mapping out such connections among every member of the group, a mathematical network representing the social relationships of that group of individuals can be generated and analyzed.

One method available to analyze such social networks is the concept of “communities” within the network, or groups of members within the larger group who are particularly closely related to each other. Thus, a group of friends who all know each other can be identified in the context of the broader network by the fact that each member of that particular group of friends shares more social connections with the other members of that group of friends than with the rest of the broader group in general.

This paper studies whether the system of judicial precedent underlying our legal system can be modeled as a social network, and whether the concept of “communities” within such a network can be used to identify groups of judicial decisions that are topically related to each other. Such a method of identifying topically related judicial decisions would offer an alternative to labor-intensive semantic analysis of the latent content of judicial decisions in order to assess their topical relationship.
In addition to potentially benefitting legal researchers seeking to identify the legal precedent relevant to a given legal issue, such techniques of network analysis may allow for novel analyses of the law leading to the development of jurisprudential theory and further elucidation of the doctrine of *stare decisis* within a particular jurisdiction. Thus, network analysis of judicial precedent might impact both practical legal research performed by practicing attorneys as well as the study and development of legal theory by law school academic researchers.

**Judicial Precedent and *Stare Decisis***

The American system of jurisprudence, like the English system at its roots, is based on the doctrine of *stare decisis*, according to which courts of law abide by their own and other courts’ prior decisions (Healy, 2001). *Stare decisis* manifests as citations within a judicial decision to prior published decisions as precedential authority for the legal principles upon which the current decision has been based.

A significant advantage of *stare decisis* is the ability to rely on the existing judicial precedent within a jurisdiction to be consistent over time, allowing one to anticipate with considerable accuracy how a court will rule on a particular legal issue in the future. However, that advantage is available only after one has identified the judicial precedent applicable to that particular legal issue. Legal research accordingly often involves the cumbersome tracing of lines of precedent among judicial decisions, in an effort to determine which prior court decisions have set out the most relevant legal principles applying to a particular set of facts. The ever-present concern in legal research is the possibility that the researcher will cease researching before encountering that one
judicial decision most relevant to a particular set of facts, thereby missing the most applicable judicial precedent.

Two tools that are particularly valuable in legal research are thus (a) the actual texts of the judicial decisions rendered in a jurisdiction, and (b) a citation index that lists the subsequent cases that cite a particular judicial decision as precedent. With those two tools, and a single judicial decision addressing a particular legal issue, it is generally possible to comprehensively explore the network of judicial precedent related to that legal issue. After exhaustively completing such exploration, a legal researcher can feel confident that he has become thoroughly familiar with the judicial precedent related to that legal issue.

However, because of the time-consuming nature and high costs associated with such an exhaustive approach to legal research, a researcher must often rely instead on secondary sources of information addressing the judicial precedent related to a particular legal issue. Secondary sources can range from treatises addressing broad areas of legal precedent, to encyclopedias of articles each addressing a narrow area of legal precedent, to topical classifications of judicial decisions based on the specific legal issues addressed therein.

The use of secondary sources of legal information can significantly shorten the time required to identify the relevant judicial precedent within a jurisdiction, but they present the inherent danger of an inadequate or incorrect understanding of the judicial precedent they address. Ideally, a secondary source would supplement, but would not replace, the thorough analysis of precedent performed directly through the primary legal decisions themselves.
Topical classification of judicial decisions.

Under the doctrine of *stare decisis*, a court must interpret the law, and address the legal issues before it, consistently with the prior judicial decisions of other courts. Unfortunately for lawyers and judges, it can be very difficult to classify the legal issues addressed in a particular judicial decision for later reference. A given judicial decision might address any number of different legal issues, any one of which could be important in regard to a subsequent legal case. Furthermore, the language used by two different courts to address related legal issues might vary greatly due to differences in the particular fact situations before each court, limiting the capacity of full text searching of prior judicial decisions to identify decisions addressing specific legal principles.

Therefore, to aid in legal research, various private companies hire lawyers to analyze the content of each new judicial decision, to summarize the key legal issues raised in that decision using standardized language, and to classify each of those legal issues pursuant to a proprietary legal classification scheme. Such legal digests are considered invaluable to legal research, but they are not infallible:

Although digests have a number of advantages, there are almost an equal number of disadvantages in their use. The digest summary may be inaccurate, may take on a different meaning in the context of the entire case, or may cover dictum rather than the holding of the case. Classification of a summary under a particular digest topic and number is an unofficial decision made by the publisher; the text of the opinion supersedes any misleading digest classification. At the same time, an important point of law from a case may have been overlooked by the publisher and lack a corresponding headnote; details important to the researcher may be overlooked. And, as noted earlier, the [West] key number system is extremely slow to react to changes in the law. (Bast and Pyle, 2001, p. 290)

As an alternative to the manual coding of latent meaning within judicial decisions, it may be possible to electronically classify such judicial decisions into communities of common legal issues by mathematically analyzing the citation network of judicial
precedent linking the judicial decisions together. Because such network community structures would be identified by means of the manifest and objective information inherent in the judicial decisions themselves, they would be free from the uncertainties of interpretation by individual content analyzers, and there would be no question of hidden bias in their classification due to the use of particular pre-existing legal classification schemes.

Accordingly, the research described in this paper addresses the question of whether algorithms designed to group network nodes by the inherent community structure of the network might be used to identify and isolate judicial decisions that address particular legal issues, based not on the latent content contained within the text of the judicial decisions but instead on the manifest structure of the decisions in the form of the citation links that connect and relate the judicial decisions with one another. The proposed research has implications across the legal field, both in the context of academic research as well as in the context of the practice of law.

**Network Analysis of Judicial Precedent**

By treating the links of precedent connecting judicial decisions as elements of a network, the current tools of network analysis may provide valuable information about the inherent relationships among judicial decisions without the need to rely on third-party analyses of precedent in the form of secondary legal sources.

**The validity of judicial precedent.**

The value of a network of judicial precedent depends in large part on the validity of the underlying citations to precedent. If a citation to precedent between two judicial decisions does not actually represent reliance by the later decision upon the legal
conclusions of the prior decision, the existence of that particular link cannot be said to inherently contain information relating the two decisions. In such a situation, information gathered from studying a network of such links would have little validity.

In an effort to determine whether citations to judicial precedent actually represent reliance on the authority of those prior decisions, Walsh (1997) studied the usage of citations among and between state courts in the context of the development in the 1980’s of the law of wrongful discharge from employment. Specifically, Walsh sought to determine whether such judicial citations “indicate intercourt communication and influence on judicial decisionmaking,” or whether they were instead “used not so much to explain the basis for decisions as to justify those decisions, however well or ill considered they may have been” (p. 339). Examining the citation patterns among 157 judicial decisions from all 50 states and the District of Columbia, Walsh found strong parallels between the courts’ analyses of the legal issues, their citations to precedent, and the conclusions they ultimately reached. Walsh concluded that the patterns he observed were inconsistent with the usage of citations to precedent merely for the purpose of justifying a decision already reached on another basis. Walsh accordingly reasoned that such citations could be relied upon as being indicative of actual lines of authority between the respective judicial decisions.

If Walsh’s conclusions are correct, it follows that analysis of the citations to precedent among judicial decisions may reliably yield substantive information about the nature of the law underlying those decisions.
Citation analysis.

Citation analysis is the study of the express links among a body of documents, and as such is a form of content analysis. Pursuant to the characterization of content analysis by Spurgin and Wildemuth (2009), citation analysis involves the analysis of the manifest links among documents, although it may also involve analysis of the latent context in which those links are made. The sampling units are the documents themselves, and the recording units are the individual links to other documents.

Smith (1981) explained that the most basic form of citation analysis is a simple citation count, “determining how many citations have been received by a given document or set of documents over a period of time from a particular set of citing documents” (p. 85). Although basic, a citation count can nonetheless be used as a measurement of the value of a document, and citation counts continue to be used for that purpose, for example in measuring the impact factors of different academic journals (Ashman, 2009). However, a citation count alone provides no information concerning the relationship between two documents.

As Smith (1981) explained, documents likely to be closely related may potentially be identified through either “bibliographic coupling” or “cocitation analysis”:

Two documents are bibliographically coupled if their reference lists share one or more of the same cited documents. Two documents are cocited when they are jointly cited in one or more subsequently published documents. . . . The difference is that bibliographic coupling is an association intrinsic to the documents (static), while cocitation is a linkage extrinsic to the documents, and one that is valid only so long as they continue to be cocited (dynamic). (p. 85).

According to Smith, citation counts and bibliographic coupling were the principle methods of citation analysis employed in the 1960’s, but lost favor to co-citation analysis in the 1970’s, particularly as a method for mapping scientific specialties. Smith noted
that the development of citation analysis was strongly influenced by the development of computers and by the availability of citation indices, such as those produced by the Institute for Scientific Information (ISI). Prior citation studies had involved hand-coding citations among documents, necessarily limiting their scope. In contrast, the advent of computers and citation indices allowed more modern citation studies to make use of large lists of citations in machine-readable form.

However, Smith (1981) cautioned that while a citation represents a relationship between a cited document and the citing document, that relationship does not necessarily represent valid information about the underlying documents themselves. An author might include a gratuitous citation to the work of another author in the field as an indication of respect, without reference to the work underlying the citation; or an author might include a self-serving citation to his own prior work, regardless of its relationship to the present work.

Smith identified five potentially erroneous assumptions that might underlie a particular citation analysis:

- “Citation of a document implies use of that document by the citing author” (p. 87).
- “Citation of a document (author, journal, etc.) reflects the merit (quality, significance, impact) of that document (author, journal, etc.)” (p. 87).
- “Citations are made to the best possible works” (p. 88).
- “A cited document is related in content to the citing document; if two documents are bibliographically coupled, they are related in content; and if two documents are cocited, they are related in content” (p. 89).
"All citations are equal" (p. 89).

Citation analysis of judicial precedent.

The principle of citation to judicial precedent shares much in common with academic citation practice, in theory as well as in actual practice. In fact, Ashman (2009) asserts that the legal profession’s Shepard’s Citations legal citation index was the inspiration behind the ISI science citation indices that so dramatically affected academic citation analysis research beginning in the 1960’s (Ashman, 2009, p. 113; Smith, 1981, p. 86). It logically follows that the tools of citation analysis described by Smith (1981) should be applicable, in some form, to the practice of citation to judicial precedent.

Along with the tools, however, are the caveats to citation analysis as enumerated by Smith (1981). Some of the caveats are of reduced concern in the context of judicial citations to precedent. For example, “[c]itation of a document implies use of that document by the citing author” (p.87) and “[c]itation of a document reflects the merit of that document” (p. 87) are fundamental principles of the doctrine of stare decisis. Citation to an irrelevant judicial precedent represents a failure of the authoring court in its practice of law and may subject the decision to reversal upon review by a higher court. As a result, courts make considerable effort to guard against such behavior.

“Citations are made to the best possible works” (Smith, 1981, p. 88) is likely subject to the same influences of convenience as are described by Smith in regard to academic citation practice. Stare decisis requires that a court act consistently with the applicable existing line of legal reasoning, but it does not obligate the court to cite to the most ideal exemplar of that line. To the extent that a judge is already familiar with a particular precedential decision relevant to a specific point of law, there is often limited
incentive for that judge to research the matter further in search of other examples of precedent relating to that issue that might be slightly more relevant to the facts at issue. It is sufficient under stare decisis merely to cite to an element in the line of existing precedential authority in order to invoke the entire line of authority.

However, Smith’s (1981) warning against assuming that “if two documents are bibliographically coupled, they are related in content; and if two documents are cocited, they are related in content” (p.89) is particularly applicable to the practice of citation to legal precedent. A given judicial decision often addresses multiple legal issues, each of which should be expected to include an applicable citation to precedent. Therefore, the principle of bibliographic coupling in the legal context is strongly influenced by the actual proximity of the citations within the judicial decision, as the mere fact that a particular decision cites two different prior decisions as precedent provides no assurance that those two decisions have been cited in regard to the same principle of law within that citing decision. Only where two citations to precedent occur close together within a judicial decision can one conclude that they are likely related to the same principle of law. Similarly, two unrelated judicial decisions might co-cite the same prior decision, citing different elements of that prior decision as precedent. Only where the two citations to the same precedent specify the same or very close page numbers within that same prior decision are the citations likely to address the same principle of law.

Finally, the assumption that “[a]ll citations are equal” (Smith, 1981, p. 89) is particularly weakly supported in the context of a judicial decision, as well-settled areas of law arising in a given judicial decision are likely to receive only cursory attention, while complex but comparatively minor issues in the decision may result in extensive
discussion regarding the applicable precedent. However, the established conventions of citation to judicial precedent include certain structural elements related to the context of citations to precedent, such as the availability of established flags to indicate the intended treatment of the precedential authority. Such manifest structural elements should provide assistance in assessing the context of the citation without recourse to extensive interpretation of the latent content surrounding the citation.

**Judicial precedent as a network.**

As explained above, *stare decisis* is manifested as citations to precedent: in principle, every legal conclusion making up a judicial decision should include a citation to one or more prior judicial decisions supporting that conclusion. It follows that the set of all American judicial decisions can be treated as a single large directed network, with the judicial decisions as nodes and the citations from newer decisions to older decisions as the links connecting the nodes together. Indeed, Smith (2007) found that the citation network defined by the body of all American judicial decisions was consistent with that of large academic citation networks.

Relatively little work has been done in the area of modern network analysis of judicial precedent, and most of that work that has taken place only within the last decade. Chander (2005), Cross, Smith, and Tomarchio (2008), Fowler and Jeon (2008), and Fowler, Johnson, Spriggs II, Jeon and Wahlbeck (2007) each examined precedential networks defined by the decisions of the U.S. Supreme Court, created either through the processing and analysis of full-text decisions or through reference to an existing citation index. In contrast, Cousins et al. (2005), McIntosh et al. (2005), and Shneiderman and Aris (2006) each studied a precedential network created by selecting a set of legal
decisions intended to encompass the entire body of federal judicial precedent related to one particular issue. Similarly, Walsh (1997) used more traditional tools of citation analysis to study a very narrow legal issue in a much more intensive fashion than the newer tools of modern network analysis would allow. Finally, Smith (2007) considered a network of judicial precedent spanning the entirety of American law.

Several of the studies of Supreme Court precedent (Chandler, 2005; Fowler and Jeon, 2008; and Fowler et al., 2007) looked at the relationship between legal decisions with high network centrality values or hub and authority scores with the decisions’ authority and importance in a legal context, while Cross, Smith, and Tomarchio (2008) measured the changes in a variety of network characteristics over time to study one particular period of Supreme Court history. Smith (2007) analyzed the broad network structure of the “Web of Law” (p. 310) and compared it to other types of large-scale networks. Taking a different tack, Cousins et al. (2005) and McIntosh et al. (2005) focused more on the construction of a network defined by the lines of precedent among and across different levels of federal courts. Shneiderman and Aris (2006) used a portion of the network created by Cousins et al. (2005) and McIntosh et al. (2005) as the basis for studying how a network of judicial precedent might be presented visually such that users could navigate the network and gain information from it directly.

Of those studies, Cross, Smith, and Tomarchio (2008) touched upon the theory of community structure within a network, but their interest was in changes in the characteristics of that community structure over time, and not in the particular grouping defined by that community structure. Only Smith (2007) actually addressed the potential
impact of examining the community structure of a network of legal precedent, and
even then only as a potential area for future research.

**Creation of a network of judicial precedent.**

The reduction of existing judicial precedent into a mathematically-defined
network is a non-trivial process. Although the underlying judicial decisions are public
record, their native format is paper, and private companies that provide electronic legal
databases tend to guard their efforts in converting and storing such judicial decisions in
electronic form. Furthermore, even with access to judicial decisions in electronic form,
any large body of such decisions must then be processed by computer to identify and
accurately code the citations to judicial precedent contained therein as network links
between the corresponding judicial decisions.

As an alternative to building a network of judicial precedent directly from the
texts of the underlying judicial decisions, a network might be built based on the
proprietary information contained in a legal citation index. To the extent that a particular
legal citation index represents an accurate distillation of the links of precedent among
judicial decisions, that index effectively describes the network of judicial precedent.

To assess the accuracy of such citation indices, Spriggs II and Hansford (2000)
examined the reliability of one legal citation index, Shepard’s Citations. That index
includes not only a list of all the citations to a particular legal decision as precedent, but
also annotations regarding the extent of the discussion of the cited decision in the citing
decision and a classification of the type of treatment received by the cited decision
(Overrule, Question, Limit, Criticize, Distinguish, Follow, Parallel, Explain, or
Harmonize) (p. 330). Spriggs II and Hansford tested the reliability of Shepard’s Citations
by identifying a random sample of actual judicial decisions and, by searching in a full-text database of judicial decisions, creating their own citation index for the sampled decisions. The authors found Shepard’s Citations to have perfect reliability in regard to the actual citations among citing and cited decisions, a result consistent with the nature of the citations as manifest content within the judicial decisions. The authors also found moderate to substantial inter-coder reliability between Shepard’s Citations and their own indexing concerning the classification of the treatment of the citation in the citing decision (with Kappa statistics between 0.401 and 0.907), a result consistent with the latent nature of such treatment within a decision.

**Networks of judicial precedent based on citation indices.**

Shepard’s Citations was used as the basis for the networks of judicial precedent studied in Cross, et al. (2008), Fowler et al. (2007), and Smith (2007). The authors in Fowler et al. (2007) were interested in studying the inherent structure of the network of precedent among U.S. Supreme Court decisions, and began with a list of every substantive judicial decision issued by the Supreme Court between 1791 and 2005, a total of 26,681 decisions. The authors then “Shepardized” each citation electronically, and used the results to define the network of citations to precedent among the 26,681 decisions. The network created by the authors was directional but un-weighted, as the authors declined to translate the citation annotations provided by Shepard’s Citations into relative link weights.

Cross et al. (2008) were interested in changes to the patterns of citation to precedent by the U.S. Supreme Court following the appointment of Antonin Scalia to the Supreme Court and the elevation of William Rehnquist to Chief Justice by President
Reagan. Cross et al. built a network of Supreme Court decisions linked by precedent extending from 1885 to 2005, using data drawn from Shepard’s Citations and provided directly to the authors by LexisNexis (the owner of Shepard’s Citations). Because the information provided by LexisNexis included even ministerial actions by the Supreme Court, it initially constituted a network of 298,566 nodes, which the authors reduced to 47,869 nodes by eliminating all decisions never cited by a subsequent decision in the network. The authors reasoned that, because they were looking for changes in the patterns of citations among decisions, there was no need to include un-cited decisions in the network.

Smith (2007), working together with LexisNexis, made use of the entirety of Shepard’s Citations and ambitiously created a network of virtually all U.S. judicial decisions, both state and federal, and all of the citations to precedent linking them. The network was made up of approximately four million decisions, and the author analyzed it variously as a whole and as sub-networks divided in accordance with the roughly 300 different jurisdictions recognized by Shepard’s Citations. Comparing those networks with existing studies of the network of academic literature in the field of high energy physics, Smith concluded that each of the networks shared a similar power-law-like distribution of citations.

*Networks of judicial precedent based on text processing.*

Other studies have not made use of an existing citation index to define their networks. Authors that built networks of judicial precedent from the actual texts of the underlying judicial decisions include Chandler (2005), Cousins et al. (2005), Fowler and Jeon (2008), McIntosh et al. (2005), and Walsh (1997. Chandler (2005) and Fowler and
Jeon (2008), like Fowler et al. (2007), were interested in the network structure of
Supreme Court precedent. Chandler created his network by processing a collection of the
texts of 26,000 Supreme Court decisions representing all Supreme Court decisions issued
between 1831 and 2005 as well as select decisions issued prior to 1831. Fowler and Jeon
(2008) similarly processed a collection of the texts of over 30,000 decisions issued by the
Supreme Court between 1754 and 2002. Each set of authors created their own software
to identify citations to precedent contained in the text of the decisions, and each
constructed a resulting network of approximately 200,000 to 250,000 links among the
various Supreme Court decisions.

Cousins et al. (2005) and McIntosh et al. (2005) addressed slightly different
aspects of the same project: an effort to map the network of citations to precedent across
all levels of federal courts in relation to the limited subject area of federal regulatory
takings. At the initial stages of the project, the authors made use of the legal
classification scheme and standardized topic descriptions provided by Westlaw, as well
as full-text searches of Westlaw’s body of federal judicial decisions, in an effort to
identify every judicial decision directly addressing the federal issue of regulatory takings
following the Supreme Court’s 1978 decision in *Penn Central Transportation Co. v. City
of New York*. Despite the difficulties faced by the authors in ultimately identifying the
complete set of 2000 to 3000 judicial decisions in which they were interested, the actual
coding of the citations to precedent within those decisions proved less complex: because
the texts of the judicial decisions being considered were obtained electronically through
Westlaw, each judicial decision included hypertext tags plainly denoting each citation to
precedent within its body. The processing and interpretation of the individual citations
and the construction of the resulting network of precedent was therefore considerably simplified from the similar task faced by Chandler (2005) and Fowler and Jeon (2008).

Finally, Walsh (1997), as discussed above, was interested in the narrow issue of the development of the law of wrongful discharge from employment within state courts in the 1980’s. Using a limited set of 157 decisions identified in the existing literature as relevant to the topic, Walsh was able to carefully select the citations within those decisions deemed to be most relevant to the topic of study. Specifically, Walsh built his network using only citations to precedent within the sections of the judicial decisions that directly addressed the topic of wrongful discharge, and additionally distinguished “strong” citations (which generally involved a quotation or discussion of the cited decision) from more rudimentary “weak” citations in his analysis of the network. However, Walsh’s meticulous analysis of the citations to precedent that he was interested in was only possible due to the extremely limited size of the network he was studying.

Network analysis of citations to judicial precedent.

Fowler and Jeon (2008) described building a network of citations among every majority opinion of the Supreme Court issued between 1754 and 2002. As noted above, the authors built their network by acquiring a corpus of 30,288 full-text majority decisions by the Supreme Court, and processing that corpus on a computer to extract 220,500 citations among those decisions. Applying the tools of network analysis to that network of citations, the authors attempted to use the network to identify “important” decisions (p. 20), and then compare their results to various expert opinions as to landmark Supreme Court cases. To identify such “important” decisions, Fowler and Jeon first considered and rejected “degree centrality,” a simple citation count, because “it treats all
inward citations in exactly the same way” (p. 20). The authors also rejected “eigenvector centrality,” “which supposes that the importance of a case is proportional to instead of equal to the importance of the cases that cite it” (p. 20), because eigenvector centrality is likewise ultimately based on a count of citations to a particular decision, without regard to other citations made by each citing decision.2

Instead, Fowler and Jeon (2008) chose to measure importance using a hub and authority model arising from social network theory. The hub and authority model determines the quality of an “authority” in a network based on the quality of the “hubs” that cite to it, and assesses the quality of those “hubs” based on the quality of the various “authorities” cited by the hub (p. 20). The result is a model that may be used to assess the importance of a Supreme Court decision based not just on the number but also on the quality of the decisions that cite to it.

Comparing the decisions characterized as important pursuant to the hub and authority model with independent expert assessments of landmark Supreme Court decisions, Fowler and Jeon (2008) found that the outcomes of the hub and authority method compared favorably with the independent expert assessments. The authors compared overall lists of the most important decisions, as well as lists of the most importance decisions in particular areas of law, as categorized by a third-party source.

Although Fowler and Jeon (2008) relied on indiscriminate computer processing to identify citations among the Supreme Court decisions they were studying, they reduced the impact of some of the limiting assumptions described by Smith (1981) by virtue of their use of the hub and authority model. Under that model, all citations are not assumed to be equal, but instead are influenced by the decisions citing and being cited. A
judicial decision with poor citations that lack relevance or merit would presumably be inconsistent with and therefore unsupported by the network of decisions around it, with a corresponding reduction in its authority value. As a result, some of the blind assumptions warned against by Smith (1981), while not eliminated, are at least counterweighted by feedback from the network as a whole.

However, to the extent that the hub and authority method rests on an implicit assumption that bibliographically coupled and co-cited judicial decisions are related, the model may not be well suited to citation analysis in general, and to the citation analysis of judicial precedent in particular, where such an assumption cannot be assumed to hold (Smith, 1981). Furthermore, the hub and authority model utilizes only a single, universal measure of authority, and therefore does not account for the particular areas of law in which a particular judicial decision might be authoritative. That shortcoming is highlighted by the reliance by Fowler and Jeon (2008) on a third-party source for the a posteriori classification of decisions by area of law in the analysis of their results.

**Community Structure in Networks of Judicial Precedent**

Once a network of judicial precedent has been compiled into one place, it can be processed by means of the appropriate algorithms for the identification of community structure within a network.

**Identifying community structure in networks.**

A network can be defined as a set of nodes joined together by links. “Examples include social networks such as acquaintance networks and collaboration networks, technological networks such as the Internet and the Worldwide Web, and biological networks such as neural networks, food webs, and metabolic networks” (Girvan and
Networks may be sparse (where the number of links closely approximates the number of nodes) or dense, with sizes ranging from a handful of nodes to hundreds of millions. The links between nodes may be directional (such as links between web pages) or non-directional, and may be individually weighted (such as by the extent of collaboration between researchers) or un-weighted.

In Girvan and Newman (2002), the authors explained that one property shared by many networks is the property of “community structure” (p. 7821). Examples of community structure within a network would include social groupings in a social network, related papers in a citation network, functional groupings in a metabolic network, or pages on a related topic on the web. The authors noted, however, that there are numerous shortfalls to traditional methods of hierarchically identifying community structure, which tended to focus on first identifying nodes with the closest associations and then working outwards to identify communities.

Instead, Girvan and Newman proposed a new method (the GN method) for identifying a network’s community structure in which the links between nodes with the highest “betweenness” (p. 7822) are recursively identified and eliminated, thereby identifying community structure from the outside in. Comparing their method with traditional hierarchical methods, the authors found their method to be significantly more effective at identifying community structure within both artificial and real-world networks. However, the authors noted limitations of the GN method, including that it assumes an un-weighted, non-directional network; that it may not perform well on dense networks with many links between individual nodes; and that it is computationally intensive, making the method unsuited to large networks.
In Newman and Girvan (2004), the same authors expanded upon their method for identifying community structure, distinguishing the GN method from graph partitioning (in which a network is divided into a fixed number of artificial communities of substantially the same size, generally for the purposes of improving parallel computer processing) and, as in 2002, from traditional agglomerative or outward-travelling hierarchical methods for identifying community structure. The authors further distinguished their method from existing methods of divisive or inward-travelling hierarchical identification of community structure, in that their method re-calculates betweenness at each recursive step, significantly improving performance over prior divisive methods but resulting in the acknowledged computationally intensive nature of their method.

In their 2004 article, Newman and Girvan also expanded upon the definition of betweenness as used in the GN method, describing various alternative algorithms for its calculation including “shortest-path betweenness,” “random-walk betweenness,” and “current-flow betweenness” (p. 3) and looking at the effects of those different measures of betweenness on the performance of the GN method. In addition, the authors defined a new unit of measure, “modularity” (p. 7), as a tool to identify the effectiveness of a particular division of a network into potential sub-communities.

Applying the GN method to additional types of networks, Newman and Girvan (2004) noted its effectiveness in identifying community structure even in a directional network such as a series of web pages, despite the fact that their method treats the links as non-directional. However, the authors once again noted the computationally intensive nature of their method ($O(n^3)$ on sparse networks), and concluded that even with a sparse
network and using the fastest method of calculating betweenness, the GN method is practical only for networks of no more than about 10,000 nodes.

In Newman (2004a), the author proposed a new method for identifying community structure in a network that bypasses the measurement of betweenness and instead makes direct use of the change in the modularity of the network, the measure introduced in Newman and Girvan (2004). Although the new method was agglomerative instead of divisive, the author found that it nearly matched the GN method in effectiveness at identifying communities within a network. The most appealing feature of the new method, however, was that it is significantly faster than the GN method ($O(n^2)$ on sparse networks), making it much more suitable for large networks. The author found that the new method could analyze a network of 1275 jazz musicians in approximately one second, while the GN method took more than three hours to reach similar results; the new method could process a network of 56,276 research scientists in 42 minutes, while the author estimated that the GN method would have taken between three and five years to complete the same calculations. Nonetheless, the author concluded that, with its slightly better effectiveness in identifying community structures, the GN method remained preferable for networks small enough to make use of that method practical.

In Clauset, Newman, and Moore (2004), the authors further refined the method described in Newman (2004a) to the point that it has “nearly linear running time” ($O(n \log^2 n)$ on sparse networks). The authors demonstrated that the refined method was effective in analyzing a network of recommender system data from Amazon.com, made up of 400,000 nodes and more than 2 million links.
More recently, Blondel, Guillaume, Lambiotte, and Lefebvre (2008) have taken that notion of a method for identifying community structure based on modularity even further, tailoring it for very large networks. Using their method, Blondel et al. were able to analyze a network of 39 million web pages and 783 million links in only 12 minutes, and an even larger network of 118 million web pages and 1 billion links was processed in just 152 minutes.

**Identification of overlapping communities.**

One potential disadvantage of each of these methods, however, is their presumption that a node within a network can belong to only one community. Thus, a node that is equally related to multiple communities is considered to belong to none of those communities individually, but instead is treated as a member only of the broader community that spans and includes each of the lesser communities – a result that arguably does not accurately reflect the involvement of that node in each of the lesser communities. As Palla, Derényi, Farkas, and Viscek (2005) noted, “most real networks are characterized by well-defined statistics of overlapping and nested communities” (p. 814). The solution proposed by Palla et al. is a method of clique percolation, making use of $k$-cliques and threshold link weights to identify communities within a network in which a single node might belong to multiple communities. Since Palla et al. (2005), various additional methods of identifying overlapping communities have been proposed.

In an effort to bridge the divide between large number of “hard-partition” community identification techniques such as the GN method and the algorithm described by Blondel et al. (2008), and the smaller body of methods to detect overlapping communities such as that of Palla et al. (2005), Wang et al. (2009) proposed a method by
which existing hard-partition techniques for identifying communities might be used to recognize overlapping communities. By introducing a local community definition with a tunable parameter $r$, Wang et al. demonstrated how the border nodes of hard-partitioned communities could be evaluated in the context of their local community and, where appropriate, assigned a divided weight representing membership across the hard-partition border. As the authors explained,

the method does not really use new ingredients or procedures, but just adjusts the classical algorithms to match the requirement for discovering overlapping communities in complex networks. However, this adjustment can lead to a considerable improvement in the community partition, like a bridge between the hard partition and the soft partition. (Wang et al., 2009, p. 5050)

Finally, because it focuses on the border nodes of an existing partition, the algorithm described by Wang et al. has the benefit of being fast, operating in nearly linear runtime ($O(n)$) on a sparse network.

**Community structure in citation networks.**

An example of the application of the GN method in identifying community structure in an academic citation network can be found in McCain (2008). In that study, the author examined the influence of the oeuvre of Conrad Hal Waddington (1905-1975). Creating a network from the set of books and articles published by Waddington, together with all of the articles identified by the Web of Science as citing those materials, and then imposing citation thresholds to limit the number of nodes to more manageable levels, the author applied the GN method to the appropriate time-framed networks of academic citations. The results provided potentially useful windows into the areas of Waddington’s work popular during different time periods, and demonstrated the
evolution of Waddington’s influence in a number of different fields over the entire 60-year period being studied.

In seeking to identify community structure within large academic co-citation networks, Wallace, Gingras, and Duhon (2009) rejected the GN method, which they acknowledged to be “the canonical method for community detection in complex networks” (p. 242), in favor of the method described by Blondel et al. (2008). In addition to its significantly improved speed over the GN method, Wallace et al. suggested that the method of Blondel et al. is more effective in identifying community structure in dense networks than the GN method. Wallace et al. applied the Blondel et al. method to a number of different author co-citation networks of varying sizes, and concluded that the Blondel et al. method was highly effective at identifying the community structure of large and complex academic co-citation networks.

Wallace et al. did note the potential drawback shared by both the Blondel et al. and the GN methods concerning the exclusivity of the community structures identified by those methods. However, while the authors acknowledged the existence of the method proposed by Palla et al. (2005) for identifying overlapping communities, Wallace et al. rejected the method due to its reliance on $k$-cliques and link weight thresholds, *a priori* factors that Wallace et al. worried might unduly confound the identification of the true inherent community structure of the networks being studied.

**Identification of community structure in citation networks.**

McCain (2008), relying heavily on the ISI citation index for scientific publications, collected citations to the work of Hal Waddington between 1945 and 2004, and then separated the citations by decade. Using UCINET software with its NetDraw
graphic display component, McCain grouped the citations within a particular decade into communities using the GN algorithm (Girvan and Newman, 2002) as implemented in NetDraw. By examining the changes over time in the communities citing Waddington’s widely varied oeuvre, McCain was able to map the evolution of Waddington’s influence within those various communities over time.

Waddington’s oeuvre seems to be particularly well suited to an historical analysis such as McCain’s, due to Waddington’s wide impact on many varied and otherwise distinct academic communities:

Not only did he publish in geology and many related areas in the biological sciences (experimental embryology and developmental biology, developmental genetics, evolutionary genetics, theoretical biology), but he also wrote on operations research in WWII . . . , ethics . . . and on the relationship between science and modern art. (McCain, 2008, p. 511)

McCain’s technique therefore might prove less effective if applied to an academic author whose citations arose primarily from a single field, since the significant changes in community over time as observed by McCain would presumably be largely absent in such a case. In addition, insofar as McCain relied on citations from ISI’s citation indexes without any further review of the primary articles, she lacked a means to ensure that the various limitations to citation analysis as described by Smith (1981) did not unduly influence her results.

Finally, McCain’s results may be influenced by sampling issues: due to the limitations of the software she was using and the GN algorithm for identifying communities, as well as an interest in simplifying the results overall, McCain only analyzed the citations among works above a certain citation threshold. McCain stated only that she selected an appropriate threshold “that struck a useful balance between
detail and clarity” (McCain, 2008, p. 512). McCain did not explore how the selected threshold may have affected her results, although she did note that “[s]ingletons (works that are cited above the threshold, but not connected to any other work) and networked pairs represent nascent networks that might (but are not guaranteed to) emerge were the citation threshold to be lowered” (p. 516).

**Community structure in networks of citations to judicial precedent.**

The goal of the present research project is therefore to blend elements of Fowler and Jeon (2008) and McCain (2008), in the hope that the more successful elements of each will be reinforced while the less successful elements will be reduced. In particular, the hope is that the community structure identification techniques demonstrated in McCain can be applied to a network of judicial precedent as in Fowler and Jeon, providing a means for the network of judicial precedent to more accurately model disparate legal subjects while retaining elements of the robustness demonstrated by the network created by Fowler and Jeon.

**Research Question and Hypotheses**

The present research project addresses the following question and hypotheses:

To what extent can topically-related judicial decisions be grouped together based on the underlying community structure of the network of judicial precedent relating the judicial decisions to each other?

Hypothesis 1: The network of citations to judicial precedent among a body of judicial decisions can be meaningfully divided into communities of decisions within that network.
Hypothesis 2: Individual communities within that network of judicial precedent will be topically related.

**Method**

**Collection of Data**

The initial plan for this project was to create a network of judicial precedent spanning the entirety of North Carolina appellate case law.

**North Carolina appellate jurisprudence.**

The North Carolina appellate court system is comprised of two court levels: the North Carolina Court of Appeals and the Supreme Court of North Carolina. The Supreme Court was established in 1799, and was the state’s sole appellate court until the creation of the Court of Appeals in 1967.

The North Carolina appellate courts rely principally upon prior North Carolina cases for legal precedent. With the exception of matters implicating federal constitutional rights (where the U.S. Supreme Court controls the law), the Supreme Court of North Carolina largely defers only to its own prior decisions in establishing new law. The North Carolina Court of Appeals is further restricted by legal precedent, in that it is obligated to follow all prior decisions of both the Supreme Court of North Carolina and of the Court of Appeals itself. While either the Court of Appeals or the Supreme Court may cite to courts of other jurisdictions for legal precedent, such citations carry no more than persuasive (as opposed to mandatory) authority.

It follows that the body of all North Carolina appellate caselaw should form a substantially closed ecosystem, wherein nearly all citations to legal precedent lie within
that system, and the few citations to external legal precedent would carry little jurisprudential import and could be disregarded with relatively little consequence.

**Sources of data.**

As discussed above, the most useful sources of data with which to create a network of judicial citations to precedent would be (a) a commercial legal citation index or (b) a comprehensive set of full-text judicial decisions.

**Commercial legal citation index.**

Shepard’s Citations, first published in 1873, is the oldest and most well-known commercial legal citation index. As a comprehensive legal citation index, Shepard’s Citations provides a list of all of the judicial decisions that have cited a particular prior decision, including information regarding the contextual treatment of the citation to that prior decision. Furthermore, as Spriggs II and Hansford (2000) showed, the citation and contextual information maintained by Shepard’s Citations can be considered to be reliable.

Creating a network of judicial precedent from the body of Shepard’s Citations data would be straightforward, as demonstrated in Cross, et al. (2008), Fowler et al. (2007), and Smith (2007). In fact, the creation of such a network would be largely trivial, as Shepard’s Citations has already grouped the data into nodes (each particular judicial decision) and directional edges connecting other nodes to that node (each subsequent judicial decision that has cited that particular judicial decision). In addition, Shepard’s Citations provides some information that would be useful in weighting those edges, based on the contextual treatment of the citation in the subsequent judicial decision.
However, reliance on the information provided by Shepard’s Citations regarding the contextual treatment of citations would ultimately limit a network based on that data, as a researcher has no ability to revise or refine the assessments of such contextual treatment as provided by Shepard’s Citations. Moreover, insofar as the contextual citation information is based on interpretation of the latent context of the citation by unknown employees of Shepard’s Citations, any network depending on such contextual information cannot be said to be based solely on the manifest structure of the underlying citation network, an important element of the research described in this paper.

Ultimately, proprietary data from Shepard’s Citations was not used as the basis for this research because it could not be obtained directly from LexisNexis, the owner of Shepard’s Citations. Although the University of North Carolina has a site license for the use of Shepard’s Citations though a web interface, that interface is designed for manual use only, and is therefore inadequate for the current research project. Therefore, this paper need not address the potential concerns raised above regarding a network based on such data.

**Published judicial decisions.**

Appellate judicial decisions are officially published in case reporters, and those printed volumes are therefore the most authoritative source for the text of judicial decisions. Furthermore, insofar as the case reporter is an official publication of a government entity, it is not subject to the same restrictions of access and usage associated with proprietary commercial data. However, because the information is in physical, printed form, it must be converted into an electronic format before it can be
systematically processed. Such a conversion on a large enough scale to meet the needs of the current research project would be an expensive and time-consuming process.

Fortunately, the Law Library Microform Consortium (LLMC) has already done much of the work to convert the printed decisions relevant to this project into a usable electronic form. LLMC is a non-profit organization that collects digital and film images of case reporters for use by member law libraries. As part of its activities, LLMC has digitally scanned and made available to its member libraries online case reporter images for a number of states, and furthermore has run optical character recognition (OCR) programs on the page images to extract their text.

LLMC maintains electronic images of all pages of the first 346 volumes of North Carolina’s Supreme Court reporter (1799 to 1997) and the first 123 volumes of the state’s Court of Appeals reporter (1967 to 1996). Because the Kathrine R. Everett Law Library at the University of North Carolina is a member of LLMC, LLMC generously made available for the purposes of this project the extracted text from all of the North Carolina court reporters that it has electronically imaged, totaling more than 436,000 pages of text.

Unfortunately, due to the limitations of the automated OCR process and the errors inherent in the text pages generated by it, the attempt to extract the judicial citation information necessary for this project from the text data provided by LLMC was ultimately abandoned. The circumstances of that decision are presented in further detail in a later section of this paper.

**Electronic judicial decisions – commercial source.**

There are a number of commercial entities that offer paid access to electronic versions of judicial decisions for the purpose of legal research. Where they have been
faced with the same potential textual errors as were generated through the LLMC OCR process, those businesses have invested heavily in improved OCR techniques and manual verification of the electronic texts, to be able to ensure their customers that the electronic decisions they are offering are accurate.

While the improved electronic judicial decisions offered by such commercial entities would be of great benefit to this project, the businesses that offer such electronic decisions generally control access to those decisions very carefully. Because the commercial entities cannot claim their own copyright on the electronic judicial decisions themselves, they must protect their investment in generating the electronic form of those decisions through strict licensing terms restricting access and use of those decisions by their users. With that understanding, and in light of the failure to obtain access even to the more limited set of data represented by Shepard’s Citations, little effort was expended on behalf of this project in acquiring such commercial electronic judicial decisions.

One commercial source of electronic judicial decisions to which much of the preceding discussion does not apply is Google’s Google Scholar service. Google currently offers access to a large number of judicial decisions in electronic form, without highly restrictive licensing of access to them. However, little information is available regarding the accuracy of those texts, and Google explicitly does not warrant that they are complete or accurate. Use of those texts is a possible area of future research for this project.

**Electronic judicial decisions – public source.**

For judicial decisions published in the last two decades, a potential alternate source for electronic versions of those decisions is the courts themselves. In North
Carolina, the North Carolina Administrative Office of the Courts (AOC) has been pre-publishing judicial appellate decisions electronically since 1995. Furthermore, since 2001 the AOC has also released electronic versions of the “unpublished” decisions of the North Carolina Court of Appeals, which are not published in the official case reporter and therefore carry no weight as legal precedent.

Although the electronic decisions made available by the AOC are not “official” decisions and therefore cannot be legally relied upon as accurate, they are generated from the same initial electronic documents that are sent to be published as case reporters, and are therefore far more reliable than electronic documents generated through an OCR process. When the AOC began making judicial decisions available in 1995, it released those decisions as HTML conversions of the original WordPerfect documents, and the imperfect conversion process did not always accurately capture the formatting of the original decision, potentially resulting in loss of information from the decision. The AOC has since switched to offering decisions only in PDF format, which faithfully replicates the format of the underlying WordPerfect document.

An unfortunate and significant drawback to the use of electronic judicial decisions pre-published by the AOC is the inherent absence of case reporter pagination information within the decisions. Insofar as legal citations to precedent may be to a particular page within a decision, the absence of pagination information prevents correlation between such a citation and a particular section of the decision. As discussed below, to the extent that this project relies upon electronic judicial decisions made available through the AOC, the lack of pagination information limits one potential direction for future research.
*Parsing of Data*

This project ultimately relied on three sources of judicial decisions as a basis for creating a network of judicial precedent: electronic text of decisions generated by LLMC; electronic decisions rendered in HTML, as provided by the North Carolina Administrative Office of the Courts (AOC); and electronic decisions in PDF form, as provided by the AOC.

**Goal of parsing.**

In order to generate a network of judicial precedent, the citations to judicial precedent must first be extracted from the component judicial decisions. Therefore, each judicial decision must be electronically parsed to identify and capture the citations to precedent within that decision. In addition, the publication citation information for the judicial decision being parsed must itself be identified, so that the network can reflect citations to that decision as precedent by subsequent judicial decisions.

As an initial stage, this research project is concerned only with the citation from one decision to another. Therefore, the parsing is limited to identifying the first instance of any particular citation to precedent within a decision, without regard to the location of that citation within the decision or to the possibility of additional citations to that same precedent within the decision. The result is an unweighted network of legal precedent between individual judicial decisions.

**Parsing efforts.**

The parsing of individual judicial decisions is simplified considerably by the existence of strict style standards for the format of citations, consisting primarily of a volume number, a reference to the court reporter being cited, and the page number for the
beginning of that decision within that court reporter volume. Additional style standards regarding the method of citing the names of the parties to the case, the year the decision was published, and where appropriate, the specific page number of the decision being cited as precedent and any subsequent appellate treatment of that decision, were not relied upon at this initial stage of the research project but may be incorporated into future research.

**LLMC.**

Because the electronic information made available by LLMC was in the form of individual pages of text from a particular court reporter, the initial step in parsing that data was to identify where each individual judicial decision began and ended within each court reporter. Unfortunately, the OCR process used by LLMC preserved little of the layout of the underlying page, making it difficult or impossible to parse the index of the case reporter to obtain information about the pagination of decisions within that reporter. Fortunately, because the individual pages within the case reporters had consistent page header information reflecting the citation information for the decision printed on that page, it was possible to parse the text of the court reporters by relying on the page headers, as well as textual indications within a page suggesting the beginning or end of a judicial decision.

The parsing of the text information provided by LLMC was performed by means of a Perl script that used regular expressions (regexp) to recognize the appropriate patterns in the text that indicated the beginning, middle, and end of a judicial decision. When the Perl script encountered errors introduced by the OCR process that it could not
process, edits were made either to the source text file or, where the error appeared to be systematic, to the actual regexp to allow the parsing to continue.

As the number of errors that had to be accounted for in the text increased, it became a concern that similar errors would have a significant negative impact on the parsing of citations to precedent within the LLMC text of judicial decisions. Errors in the page headers and in the text that began or ended a judicial decision could be recognized and compensated for because that information was relied upon by the parser, and in its absence parsing could not continue. But similar errors in the citations to precedent within a decision would be very difficult to identify, and therefore much more difficult to compensate for. A parser that collected only the bare minimum of information from a citation – the volume, reporter, and page number – would not even recognize a changed volume or page number as an error, but would instead simply collect erroneous data. Furthermore, an error that caused a citation to diverge too significantly from the standard style of citations would simply be skipped by the parser as non-citation text. It follows that code to adequately parse the citations to precedent within the LLMC judicial decisions would need to be robust enough to identify a wide range of sequences of text that might be citations to precedent, as well as to rely on additional citation information such as party names and decision year to identify and confirm citations to a particular prior judicial decision.

Creating a sufficiently robust parser to adequately parse the LLMC judicial decisions exceeded the scope of this research project. Therefore, the process of parsing the LLMC text was terminated after extracting just 1,515 decisions from thirteen volumes.
of the case reporter for the North Carolina Court of Appeals, spanning years 1993 to 1996.

**AOC HTML.**

After abandoning the parsing of the LLMC text, the research project focused on the more limited set of North Carolina appellate judicial decisions made available by the AOC. A set of electronic judicial decisions in HTML format was initially parsed to collect the information sufficient to create a pilot network of judicial precedent.

The decision to parse HTML decisions was based largely on the ease of parsing the HTML code, insofar as extracting the underlying text of the judicial decision is simply a matter of removing the tagged HTML elements from the document. The pilot network includes citation information from unpublished decisions where they are available, since they represent valid nodes in a network of citations to precedent, albeit nodes with no inbound citations possible.

The pilot network is based on a data set of 13,322 decisions of the North Carolina Court of Appeals issued between 1995 and 2008, including unpublished decisions issued between 2001 and 2008. In 5,074 of those decisions, the HTML code of the judicial decision had been tagged with the citation information for the decision. The remaining decisions were either unpublished decisions or were decisions that had not yet been updated with official citation information at the time they were collected for this project.

Of those 13,322 decisions, 43 were rejected as duplicative, since they included citation information that matched another judicial decision in the data set that had already been processed. Such a situation can arise where a single judicial decision addresses multiple appellate cases, and the AOC provides a separate but identical HTML document
for each underlying appellate case. An additional 147 decisions did not include any
identifiable citations to judicial precedent, and were not cited by any other decisions in
the data set as precedent, and so were eliminated as isolated notes within the network.

The remaining 13,132 electronic judicial decisions were combined with 20,207
inferred judicial decisions not included in the data set but cited by at least one of the
electronic judicial decisions in the data set. The inferred judicial decisions include
decisions of the North Carolina Court of Appeals issued before 1995, as well as decisions
of the Supreme Court of North Carolina, and are included in the pilot network to allow
for the co-citation of a common inferred decision by multiple judicial decisions within
the data set. Those 33,339 judicial decisions are represented as nodes in the pilot
network of citations to judicial precedent, and are linked by 135,972 edges representing
individual citations to precedent between two nodes.

**AOC PDF.**

After the generation of the pilot network, a more robust network of citations to
judicial precedent was created based on the electronic judicial decisions provided by the
AOC in PDF format. That data set of appellate judicial decisions includes most decisions
issued from 1998 through July 2011 by both the Supreme Court of North Carolina and the
North Carolina Court of Appeals, including unpublished decisions of the Court of
Appeals beginning in 2002. The data set is made up of a total of 15,913 judicial
decisions, including 1,245 decisions of the Supreme Court of North Carolina, 7,393
published decisions of the Court of Appeals, and 7,275 unpublished decisions of the
Court of Appeals.
The inclusion of Supreme Court decisions in the data set has a potentially significant impact on the network of judicial precedent generated from the data set. Although the total number of Supreme Court decisions is small compared with the number of Court of Appeals decisions, the legal precedent established by the Supreme Court decisions is binding on the Court of Appeals. As a result, the Supreme Court decisions are likely to be highly cited by the Court of Appeals decisions, and therefore to play a significant role in the network of judicial precedent generated from the complete data set.

A benefit of the PDF decisions over the HTML decisions is their faithful reproduction of the layout and content of the original electronic documents generated by the Courts. However, a consequence of that nature is that the PDF decisions, unlike the HTML decisions, are not updated with case reporter citation information once a decision has been published. Instead, citation information for the individual decisions is maintained in the index to the decisions provided by the AOC. Therefore, parsing the PDF decisions first requires parsing the indices to the decisions provided by the AOC.

The requirement to parse the indices to the PDF decisions brings several advantages for the research project. Because the indices group citation information for decisions by the date the decisions were issued, the indices present an opportunity to identify data errors though inconsistencies among groups of citations that should be similar. As a result, the network of judicial precedent based on the PDF decisions includes a level of data cleanup of the individual decision citation information that was not possible in the pilot network generated solely from the HTML decisions provided by the AOC.
In addition, for decisions issued after 2000, the indices provide extremely limited topical information for most published decisions. The topical information is subjective and not based on any classification scheme, but the content information it provides offers an alternative to more extensive latent content analysis of individual decisions to assess topical relevance in the analysis component of this research project.

Unfortunately, a significant drawback to parsing judicial decisions in PDF format is the difficulty in extracting the actual text of the judicial decision from the PDF document. For this project, the text was extracted using the open source iText Java library for PDF manipulation, and custom Java code had to be created to deal with the presence of WordPerfect-specific fonts used to represent punctuation in many of the PDF documents. Furthermore, 654 of the PDF decisions in the data set included some amount of text that the iText library could not extract. Because all but three of those decisions were issued after February 2011, the problem appears to be due to an internal shift around that time by the North Carolina appellate courts from WordPerfect to Microsoft Word as their word processor software, and the tendency of Microsoft Word to occasionally encode English text in a manner normally reserved for encoding extended character sets such as those used in Asian languages.

Of the 15,259 PDF decisions from which the text of the decision could be extracted, 49 decisions were rejected as duplicates of other decisions in the data set. Of the remaining decisions, which were processed through a Perl script similar to the one used for the AOC HTML decisions, 231 decisions had no citations of precedent and were not cited as precedent by any other decision in the data set (many of which are likely summary “per curiam” opinions issued by the Supreme Court which, in lieu of citations
to precedent, refer only the Court of Appeals decision being appealed). 20,534 additional decisions were inferred from citations by decisions in the data set. The result is a network with a total of 35,513 nodes representing actual and inferred decisions, and 167,627 edges representing unique citations to precedent between those decisions.

That network of citations to judicial precedent spans only 7% more judicial decisions (actual or inferred) than the pilot network, but includes 23% more citations to precedent among those decisions. The significant increase in the number of citations to precedent represented by the network is plausibly due to the inclusion of Supreme Court decisions in the data set.

Detection of Communities

Once the network of judicial precedent has been built, it must be processed through a community detection algorithm in order to generate the data to be analyzed in accordance with this research project.

Perl implementation.

The initial community detection algorithm selected for this project was the modularity-based algorithm described in Newman (2004a). While that algorithm may not be quite as effective at identifying community structure as the GN method described in Girvan and Newman (2002), it operates considerably more quickly than the GN method (O(n^2) instead of O(n^3) on sparse networks) and has the added advantage of inherently identifying the point at which the division of the network into communities has reached maximum modularity, and thus the point at which the communities identified in the network have been most clearly defined.
An implementation of the Newman (2004a) algorithm was written in Perl, and tested against a series of small networks to confirm its effectiveness in identifying communities within a network. However, an attempt to process the pilot network through the Perl implementation of the Newman (2004a) algorithm proved prohibitively slow on a home computer.

Therefore, a sub-network of the pilot network was created by identifying all of the AOC HTML decisions that include the phrase “Industrial Commission,” a quasi-judicial State agency from which appeals may be taken directly to the North Carolina Court of Appeals. Of the 13,132 AOC HTML decisions making up the pilot network, 991 include the phrase “Industrial Commission,” or 7.5% of the total set of decisions. The sub-network of judicial precedent based on those 991 decisions is comprised of a total of 3,674 nodes, including inferred judicial decisions, and 11,185 edges representing individual citations to precedent. Thus, the sub-network consists of 11% of the total nodes of the pilot network and 8.2% of the citations to precedent present in the pilot network.

The Perl implementation of the Newman (2004a) algorithm took nearly three days to process the sub-network on a home computer. Even on the Kure research computing cluster at the University of North Carolina, the Perl script took more than 44 hours to complete. At $O(n^2)$, a network ten times larger than the sub-network would be expected to take 100 times longer to process. It follows that the Perl implementation of the Newman (2004a) algorithm created for this project cannot practically be used on networks much larger than the sub-network.
**R/igraph library.**

The open-source igraph library of network analysis and visualization functions, which is available as a package for the statistical programming language R, includes an implementation of the Clauset, Newman, and Moore (2004) community detection algorithm. Clauset et al. (2004) describe their algorithm as a mathematical refinement of the Newman (2004a) algorithm, returning the same results as the Newman (2004a) algorithm but operating in nearly linear running time (O(n log^2 n) on a sparse network). As with the Perl implementation of the Newman (2004a) algorithm, the igraph implementation of the Clauset et al. (2004) algorithm was successfully tested against a series of small networks to confirm its effectiveness. However, in striking contrast to the processing speed of the Perl implementation of the Newman (2004a) algorithm, the igraph implementation of the Clauset et al. (2004) algorithm completed processing the 3,674-node sub-network of the pilot network in less than one second.

An attempt to process the entire pilot network through R/igraph failed because igraph rejected the pilot network as not being a “simple graph.” The most likely cause of a network not being a “simple graph” would be the presence of self-citations and circular citations. Because self-citations do not represent citations to precedent, but instead are citations to the decision itself, they should not be included in a network of citations to precedent. Circular citations arise from two decisions citing each other, and can therefore be reduced to a single edge between the two decisions in an undirected and unweighted network. The parser for the AOC PDF decisions was written to prevent self-citations and circular citations from appearing in the network being generated from those decisions.
The 35,513-node network generated from the AOC PDF decisions was successfully processed by means of the igraph implementation of the Clauset et al. (2004) community detection algorithm in under two minutes.

Results

The results of processing the networks of citations to judicial precedent through community detection algorithms are as follows.

Pilot Network

As noted above, only the sub-network of the pilot network that includes judicial decisions containing the phrase “Industrial Commission” was processed through the community detection algorithms studied in this paper. That sub-network was processed through both the Perl implementation of the Newman (2004a) algorithm and the igraph implementation of the Clauset, Newman, and Moore (2004) algorithm.

Communities generated.

The Perl implementation of the Newman (2004a) community detection algorithm identified 175 groups in the 3,674-node sub-network of the pilot network. Of those groups, 54% of the nodes of the sub-network are associated with the largest three groups, and 85% of the nodes are associated with the largest ten groups. Of the remaining groups identified as communities, 84% of the groups (147 of 175) have five or fewer nodes, and 74% of the groups (129 of 175) contain only a single node.

The igraph implementation of the Clauset et al. (2004) algorithm identified only 47 groups in the same 3,674-node sub-network. Of those groups, 58% of the nodes are associated with the largest three groups, and 89% of the nodes are associated with the
largest 10 groups. However, only 32% of the groups (15 of 47) have five or fewer nodes, and no groups contain only a single node.

There is limited overlap among the groups identified through the Newman (2004a) algorithm and those identified through the Clauset et al. (2004) algorithm, as demonstrated in Table 1 below. For example, Table 1 shows that the second-largest group identified through the Newman (2004a) algorithm (Group II), made up of a total of 656 judicial decisions, shares 139 of those decisions with the largest group identified through the Clauset et al. (2004) algorithm (Group A) and 364 decisions with the third largest group identified through the Clauset et al. (2004) algorithm (Group C). The remaining 153 decisions that make up Group II from the Newman (2004a) algorithm are shared with smaller groups identified through the Clauset et al. (2004) algorithm.
Table 1

*Overlap among Different Groups of Judicial Decisions*

Three largest groups identified through the Clauset et al. (2004) algorithm

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>(Remaining decisions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>815</td>
<td>756</td>
<td>562</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td></td>
<td>139</td>
<td>0</td>
<td>364</td>
</tr>
<tr>
<td>III</td>
<td></td>
<td>509</td>
<td>5</td>
<td>12</td>
</tr>
</tbody>
</table>

Three largest groups identified through the Newman (2004a) algorithm

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>B</th>
<th>C</th>
<th>(Remaining decisions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>14</td>
<td>553</td>
<td>9</td>
<td>(114)</td>
</tr>
<tr>
<td>II</td>
<td>139</td>
<td>0</td>
<td>364</td>
<td>(153)</td>
</tr>
<tr>
<td>III</td>
<td>509</td>
<td>5</td>
<td>12</td>
<td>(127)</td>
</tr>
</tbody>
</table>

Full Network

The 35,513-node network of citations to judicial precedent built from the AOC PDF judicial decisions was processed through the igraph implementation of the Clauset et al. (2004) community detection algorithm.
Communities generated.

Processing the full 35,513-node network of citations to judicial precedent generated a set of 174 groups of judicial decisions. Of those groups, 83% of the nodes of the network are associated with the largest three groups, and 97% of the nodes are associated with the largest ten groups. Of the remaining groups identified as communities, 63% of the groups (109 of 175) have five or fewer nodes, and none of the groups contain only a single node.

If only the 14,979 nodes representing actual judicial decisions are considered, there are still 174 groups, with 85% of the nodes associated with the largest three groups and 98% of the nodes associated with the largest ten groups. However, 90% of the groups (157 of 174) have five or fewer nodes, and 62% of the groups (108 of 174) contain only a single node.

Community content.

Based on the names of the parties to the AOC PDF decisions that are the basis for the network of citations to precedent, as well as the limited subject information provided by the AOC for some of the decisions, certain assessments of the content of particular communities as identified by the Clauset et al. (2004) algorithm is possible.

For example, criminal cases can be identified by the plaintiff’s name of “State,” and the 6,207 AOC PDF decisions for those criminal cases span from State v. Aaron to State v. Zuniga. Those criminal cases make up 41% of the full set of 14,979 AOC PDF decisions. Table 2 shows the percentage of such criminal cases associated with each of the ten largest groups identified by the community detection algorithm. As can be seen from Table 2, the first, fifth, and tenth largest groups of judicial decisions consist almost
entirely of criminal decisions. The remaining seven largest groups include no more than 15% criminal decisions, with the majority of them below 2%.

Table 2

*Population of Criminal Decisions within the Ten Largest Communities*

<table>
<thead>
<tr>
<th>Rank of Group</th>
<th>Total Size of Group</th>
<th>Criminal Decisions</th>
<th>Percentage of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>5,291</td>
<td>5,132</td>
<td>97%</td>
</tr>
<tr>
<td>II</td>
<td>4,784</td>
<td>30</td>
<td>0.63%</td>
</tr>
<tr>
<td>III</td>
<td>2,639</td>
<td>90</td>
<td>3.4%</td>
</tr>
<tr>
<td>IV</td>
<td>857</td>
<td>117</td>
<td>14%</td>
</tr>
<tr>
<td>V</td>
<td>758</td>
<td>716</td>
<td>94%</td>
</tr>
<tr>
<td>VI</td>
<td>111</td>
<td>2</td>
<td>1.8%</td>
</tr>
<tr>
<td>VII</td>
<td>91</td>
<td>8</td>
<td>8.8%</td>
</tr>
<tr>
<td>VIII</td>
<td>63</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>IX</td>
<td>38</td>
<td>1</td>
<td>2.6%</td>
</tr>
<tr>
<td>X</td>
<td>26</td>
<td>22</td>
<td>85%</td>
</tr>
<tr>
<td>(Remaining Groups)</td>
<td>(321)</td>
<td>(89)</td>
<td></td>
</tr>
</tbody>
</table>

As an example of the information available through the limited subject information provided by the AOC for some decisions, consider the tenth-largest community described above. Of the 26 decisions that make up the community, twelve have subject information associated with them. Of those twelve elements of subject information, eleven expressly refer to bail bond forfeiture and the twelfth includes
reference to a bail bondsman. It is therefore highly likely that all 26 of the decisions making up that community are related to bail bonds and bail bond forfeiture.

**Discussion**

The results described above provide insight into the effectiveness of the application of community detection algorithms to a network of judicial citations to precedent.

**Community Detection Algorithms**

The results show that the communities generated by the Perl implementation of the Newman (2004a) algorithm do not completely align with the communities generated by the igraph implementation of the Clauset et al. (2004) algorithm. That result appears inconsistent with the statement in Clauset et al. (2004) that the algorithm described therein “performs the same greedy optimization as the algorithm of [Newman (2004a)] and therefore gives identical results for the communities found,” (p. 066111-1), albeit in a significantly more efficient manner.

An explanation for the differences in results may lie in the description of the Newman (2004a) method set forth in Clauset et al. (2004), wherein the Newman (2004a) algorithm is described in terms of matrix calculations. The Perl implementation of the Newman (2004a) algorithm does not rely on matrix calculations, but instead on a less efficient looping navigation of the network. If inefficiencies introduced in the Perl implementation of the Newman (2004a) algorithm manifest in the communities being generated by that algorithm, differences between the Perl and the igraph implementations of community detection algorithms may not be unexpected.
Alternately, it is possible that either the Perl implementation of the Newman (2004a) algorithm or the igraph implementation of the Clauset et al. (2004) algorithm is flawed, and the data generated from that algorithm is inconsistent with expectations. In such a case, it is most likely that the Perl implementation of the Newman (2004) algorithm is flawed, since that implementation was created specifically for this project. However, if there is a flaw in either implementation of the community detection algorithm, it is a subtle flaw, as both algorithms were successful in identifying community structure in small test graphs, and both appear to identify community structure to some extent on the scale of the sub-network of the pilot network of citations to judicial precedent.

**Communities within a Network of Citations to Judicial Precedent**

The results described above appear to show, in accordance with the hypotheses of this paper, that a network of citations to judicial precedent has sufficient modularity to describe communities of judicial decisions. Furthermore, there is evidence to support the conclusion that the decisions within a single community can, in at least some cases, be topically related. However, more detailed analysis of the latent content of the decisions within a given community will be needed before the topical relationship of decisions within a single community can be further assessed.

The results also appear to show a number of decisions within the network of judicial precedent that are not associated with a community of other decisions, but are isolated into unitary or very small groups. Such exceptions seem inconsistent with the hypothesis that judicial decisions can generally be assigned to communities of topically related decisions. The likelihood that each of those isolated groups of decisions

...
represents a very small topical group of decisions substantially unrelated to any other larger group of decisions seems unlikely.

A more likely explanation is that the isolated decisions are an artifact of the inclusion of inferred decisions in the network of judicial precedent. As noted in the results, 62% of the groups identified by the community detection algorithm consist of a single AOC PDF judicial decision coupled with at least one inferred decision cited by that PDF judicial decision. Given how the Newman (2004a) and Clauset et al. (2004) community detection algorithms function, it seems plausible that, where an inferred decision is cited by only a single full-text decision, the artificial pairing of the full-text decision and the isolated inferred decision could be separated out as their own community. If the inferred decision were available in full text, the network would include the citations to precedent contained in that inferred decision, preventing the inferred decision from appearing to be an isolated decision linked to the rest of the network of judicial precedent by a single citation from a single decision.

Conclusion

The research presented in this paper supports the hypothesis that the network of citations to judicial precedent among a body of judicial decisions might be used to group those decisions into communities of common topicality, based only on the objectively manifest structure of the network of citations to precedent and without the need to assess the latent content of the individual decisions. However, as discussed below, there are numerous possibilities for the refinement and expansion of this research.
Potential Impact of Research

This research has potential impact in the areas both of practical legal research and of academic legal theory.

**Legal research.**

The potential ability to identify topically related judicial decisions solely through the objective structure of a body of judicial decisions offers the opportunity for new tools for legal research, which would complement both comprehensive and exhaustive legal analysis as well as existing human-mediated secondary sources of legal research. Providing the legal researcher with information regarding the community of decisions linked to a particular judicial decision through a web of citations to precedent may help the researcher find related decisions based on the treatment of those decisions by the citing courts, even if those decisions do not align predictably with the assessment of topicality assigned to those decisions by a third-party legal digest.

**Legal theory.**

In addition, the ability to compare the communities defined by the structure of a network of judicial precedent with topical communities identified through the analysis of the latent content of those judicial decisions may offer valuable insight into the nature of judicial precedent and *stare decisis*. Differences between the topical communities identified through the manifest structure of a citation network and the topical communities identified through the analysis of latent content could provide a valuable method to study the boundaries of *stare decisis* and the limits of the relationship between judicial issues on the one hand, and the citations to precedent used by the authoring courts to provide support for the resolution of those issues on the other hand.
Improvements to Current Research

There are a number of areas in which the research presented in this paper could be improved or extended, with relatively little difficulty.

**Judicial decision data.**

The network of citations to judicial precedent generated from the AOC PDF judicial decisions for this project included some areas of data clean-up, but more data clean-up is possible.

**Indices.**

Further data clean-up could be performed on the indices providing the publication citations for the AOC PDF judicial decisions. Gaps in the publication citation information in the indices raise the possibility of citations to judicial precedent that could not be accurately linked to the prior decision present within the data set. To the extent that the body of AOC PDF judicial decisions includes every judicial decision published by the North Carolina appellate courts over a certain span of time, it follows that any precedential citation to a prior decision issued within that span of time should be associated with a decision in that body of decisions, and there should be no inferred judicial decisions within that span of time.

**Body of decisions.**

In fact, the body of AOC PDF judicial decisions does not include every judicial decision published by the North Carolina appellate courts within the relevant span of time, as reflected in the index by decisions marked “Opinion not available.” Efforts could be made to obtain copies of those decisions, wherever possible.
Conversion of PDF to text.

Finally, there is the issue of conversion of the judicial decisions into a form suitable for parsing. The conversion of PDF to text can present difficulties when elements of formatting are not adequately converted, such as the distinction of footnotes from the rest of the text on a page. In addition, as noted earlier in this paper, approximately 5% of the AOC PDF judicial decisions were rejected because of font encoding issues that could not be resolved by the process used to convert the PDF decisions to text. Improvements in the conversion of the PDF judicial decisions into text would broaden and strengthen the body of decisions underlying the network of citations to judicial precedent being generated.

Judicial decision parser.

Two areas where the judicial decision parser could be improved are the parsing of individual citations to precedent, and the handling of inferred decisions within the network of citations to precedent.

Parsing of citations.

Of the potential improvements to the judicial decision parser, the most valuable would likely be improvements in the recognition of citations to precedent. The parser created for this research relied almost exclusively on the volume and page information provided in a citation as the basis for identifying a citation to precedent. However, in an effort to track data quality, additional information about the party names associated with that citation were tracked and monitored. From that data it is apparent that, even with the existence of strict conventions for the format of citations, there are inconsistencies in the
actual citations to precedent among the judicial decisions making up the network studied for this project.

Of 16,932 actual and inferred judicial decisions cited as precedent at least twice within the body of AOC PDF decisions, 40% of those decisions (6,750) were cited inconsistently on at least one occasion. Many of those inconsistencies are minor variations in the formatting of the names of the parties to the case. But many other inconsistencies are not as clearly resolvable, and some may represent typographical errors in the volume and page information for the citation. As an improvement to the judicial decision parser, the parser could cross-check such auxiliary citation information against the index of publication citations provided by the AOC, as well as other instances of purported citations to the same precedent, as a means of identifying citation errors before they are integrated into the network of judicial precedent.

**Inferred decisions.**

Inferred decisions are included in the network of citations to judicial precedent studied in this paper because the inferred decisions offer potentially valuable co-citation information, indicating a link between two full-text decisions in the network that both cite as precedent the same prior decision not available as a full-text decision for parsing.

However, as noted in the discussion section above, it is likely that the presence of inferred decisions cited by only a single full-text decision are disrupting the community detection algorithms by incorrectly implying network termination points that the inferred decisions do not actually represent. It is therefore possible that the elimination from the network of inferred decisions cited as precedent by only a single actual decision in the
data set would improve the performance of the community detection algorithms applied to the network.

On the other hand, elimination of only single-citation inferred decisions in an effort to improve the structure of the network of citations to precedent raises the specter of selective data collection, since an entire class of citations to precedent within the set of full-text decisions would no longer be represented in the network.

**Output analysis.**

The analysis of the community structure of the networks of citations to judicial precedent studied in this project could be extended in a number of ways. In particular, further information about the nature of a community could be provided through additional analysis of the latent content of the decisions in the community, as well as analysis of the sub-community structure of the community.

**Latent content analysis.**

The analysis of individual communities identified by this project in a network of citations to judicial precedent is quite rudimentary, relying largely on citation counts, case party names, and very limited subject information available for some of the AOC PDF decisions. The project does not include any analysis of the latent content of the decisions making up a particular community, beyond the limited subject information provided by the AOC.

A number of methods exist to provide additional information regarding the latent content of the decisions in a community. One option, available only on the scale of a small community, would be to review each decision in the community individually and classify its latent content manually. Another option would be to compare the
communities found in the network with existing third-party assessments of the topicality of the individual judicial decisions within those communities.

A third option would be to collect analyze the decisions grouped into a community by collecting information about the frequency of various $n$-grams, or sequences of words of length $n$, within the decisions. By comparing the frequencies of various $n$-grams within a community with the frequencies of those same $n$-grams among the group of judicial decisions as a whole, it may be possible to infer the latent content of the decisions that is common within the community. Furthermore, it might be possible to compare those $n$-gram frequencies for the community as a whole with the $n$-gram frequencies for individual decisions within the community, in order to assess the consistency of such latent content within the community.

*Further community identification.*

This project considered only the top-level set of communities identified in a network of citations to judicial precedent. However, each of the communities identified could itself have been processed through the community identification algorithm and separated into component sub-communities. Thus, where a single large community might not have a clear common topicality among its decisions, that large community might naturally divide into smaller communities each with a more apparent common topicality among its decisions.

It is worth noting, however, that the community detection algorithms described in Newman (2004a) and Clauset et al. (2004) are based on the concept of modularity, and define the communities existing in a network as the communities identified at the maximum modularity of that network. Therefore, even if a large community does not
readily share a common topicality across all of its decisions, that community was
defined as a single community because the community detection algorithm found, on the
basis of the citations to precedent, that all of the decisions within that community had
more in common with each other than with the decisions assigned to other communities
by the algorithm.

Further Research

The following are some areas in which the research underlying this paper could be extended, although not as readily as the areas described in the immediately preceding section.

Weighted edges.

In the networks of judicial precedent studied in this paper, a citation to judicial
precedent is represented by a single unweighted edge between the two nodes representing
the citing judicial decision and the cited judicial decision. However, the treatment of a
particular citation to precedent within a given judicial decision may vary widely, and
information regarding that contextual treatment may be determined by examining the
manifest structure of that decision. Thus, a citation to a particular prior decision as
precedent might be given increased weight if (a) the text of the citing decision includes
multiple citations to that same prior decision, (b) the citing decision quotes from the prior
decision, or even (c) the citing decision cites to a specific page of the prior decision.
Conversely, a particular citation to precedent might be given reduced weight if it only
appears as one in a string of citations to precedent.

The judicial decision parser could be expanded to identify such additional
information about citations to precedent, and reflect those factors as weighted edges in
the network of judicial precedent. The community structure detection algorithm would then take the weight of those edges into account when assessing divisions between communities.

However, modifying the judicial decision parser to provide for weighted citations to precedent is not a trivial process. In particular, identifying multiple citations to the same prior decision as precedent introduces the difficulty of parsing the citation “id.”, which indicates another instance of the immediately preceding citation in the judicial decision. The problem with parsing “id.” is that it does not apply solely to citations to North Carolina appellate precedent, but may refer to a prior citation to statute, to secondary authority, or even to a judicial decision from a different jurisdiction cited as persuasive authority. Therefore, to account for multiple citations to precedent in weighting the edges of a network of judicial precedent, the decision parser must either (a) ignore the use of “id.”, potentially reducing the quality of the data being gathered, or (b) be capable of recognizing and rejecting a very wide range of possible citations, so that only instances of “id.” reflecting citations to applicable judicial precedent are reflected in the network.

In addition, it is worth noting that there may be a negative impact to weighting the edges of a network of judicial citation based on the manifest structural treatment of the citations to precedent. As noted earlier in this paper, a citation to well-established precedent may be cursory, even if it is directly relevant to the issues presented in the citing decision, simply because it is not a point in legal dispute in the decision. It is unclear what the effect would be on a network of citations to judicial precedent if a
citation to highly relevant but well-settled precedent were weighted equally with a citation with little relevance cited in passing in the citing decision.

**Nodal statutes.**

In addition to citations to precedent, judicial decisions may include citations to statutes, which are binding on courts in a manner similar to binding precedential authority. By treating statutes similarly to inferred judicial decisions, the network of judicial precedent can benefit from the information implied in the co-citation of the same statute by two different judicial decisions.

A potential downside to treating statutes similarly to precedential decisions is the changeability of statutes: if a statute is legislatively amended, a judicial decision citing the statute prior to its amendment may be less related than would otherwise be anticipated to a judicial decision citing the statute after its amendment. It is possible, therefore, that inclusions of citations to statutes in a network of judicial precedent could result in a degradation of the quality of that network.

**Divided nodes.**

A likely significant limitation to the application of community structure detection algorithms to a network of judicial precedent is the fact that legal decisions are not limited to a single subject. Because of the nature of *stare decisis*, judicial decisions must cite to legal precedent on even the most incidental of legal issues arising in a particular case. Insofar as each decision may be made up of a unique collection of legal issues, assigning a decision into a single community of judicial decisions likely involves dismissal or minimization of certain legal issues not consistent with the overall community.
One possible approach to addressing this problem might be to represent a single judicial decision as a series of related nodes in the network of judicial precedent, instead of as a single node. In that manner, the section of a decision addressing one legal issue could be grouped into a separate community from a different section of the decision addressing a distinct and unrelated legal issue.

Unfortunately, most appellate decisions are not clearly divided into separate sections for each legal issue addressed. However, most decisions do linearly address one legal issue before moving on to another issue. Therefore, by dividing a decision into a linear series of sub-nodes, citations to precedent within that decision could be assigned to a particular sub-node based on their relative location within the decision. Even if a particular legal issue addressed in a decision spans multiple sub-nodes, each sub-node will be associated with the subsequent and prior sub-nodes for that judicial decision, so citations to precedent in adjacent sub-nodes will be associated with each other through the adjacency connection between those sub-nodes.

Citations to a judicial decision as precedent present a greater difficulty for the sub-node model. Citations to precedent that do not refer to a specific page number of a particular prior decision would presumably have to be associated with all sub-nodes of that decision; furthermore, the weight assigned to the network edges representing those citations would presumably have to be attenuated, depending on the number of sub-nodes in the prior decision, to distinguish a single citation to the decision as a whole from multiple individual citations to individual sub-nodes of the same decision.

Citations to precedent that specify a particular page number present a different difficulty. If the pagination of the prior decision is known, then the decision could be
assigned one sub-node per page and inbound citations could be associated with a specific sub-node. But in situations such as the AOC PDF judicial decisions where pagination information is unavailable, a citation to a specific page of a prior decision can at best be associated with the sub-node (or sub-nodes) likely to represent that intended section of the decision, based on factors such as the length of the decision and, potentially, other citations to the decision that also specify a page number.

**Overlapping communities.**

As an alternative to dividing a judicial decision into sub-nodes, a community detection algorithm that allows for overlapping communities could be used to capture the divergent associations arising out of the presence of multiple legal issues in a particular judicial decision. Thus, a single judicial decision could potentially belong to multiple communities, one for each of the different legal issues presented in that single decision.

A problem presented by algorithms that provide for overlapping communities, as noted by Wallace, Gingras, and Duhon (2009), is the requirement for at least one additional parameter to reflect the appropriate threshold upon which communities may overlap. Because that parameter must be tuned, the algorithm necessarily becomes more subjective and less generalized than, for example, a community detection algorithm based solely on the objective modularity of a network as a whole.

**Expanded network.**

The research presented in this paper is significantly restricted by the limited set of judicial decisions available from which to generate the network of citations to judicial precedent. The network built from the AOC PDF judicial decisions includes more decisions inferred from citations to precedent than actual decisions in the data set,
resulting in an incomplete network with the inferred decisions representing artificial network termination points. As discussed above, the large number of small independent communities identified by the community detection algorithms used in this project are likely due to the presence of the inferred decisions, and the absence of citation information for those inferred decisions that would link them to yet decisions within the network.

As originally conceived, this research project had the goal of creating a network of citations to precedent spanning the entirety of North Carolina appellate jurisprudence over a period of more than 200 years. Because the vast majority of the judicial precedent relied upon by those decisions would fall within that network, the need for inferred decisions would be significant reduced if not entirely absent. The result would likely be a network for which the application of community detection algorithms would be significantly more effective.

Unfortunately, as discussed in some detail above, obtaining the citation information necessary to create such a complete network of citations to judicial precedent presents numerous difficulties. Access to commercial sources of judicial case information is generally extremely restricted. However, alternatives to such commercial sources, such as the LLMC data and, most likely, Google Scholar, present data quality issues that are likely to significantly impede the development of an accurate network of judicial precedent.

Nonetheless, acquiring such additional judicial case information and expanding the citation network accordingly would be among the most effective means of extending and validating the research presented in this paper.
NOTES

1 Alternately, a legal conclusion might be directly supported by statute, which can be thought of as the legislative process of short-circuiting *stare decisis*.

2 Google’s PageRank algorithm for prioritizing web pages is a variant on eigenvector centrality.

3 In Newman (2004b), the author demonstrated that the GN method for identifying community structure can be applied to networks with weighted links or with multiple links between nodes, by mapping such networks onto multigraphs and then applying the same techniques to the multigraphs as would otherwise be applied directly to an unweighted network.

4 An incidental benefit to the use of “modularity” as the determinative factor is the new method’s inherent capacity to analyze weighted networks.

5 LLMC was a significant source for the electronic judicial decisions now offered by Google Scholar.

6 Because the pilot network and its sub-network are not actually sparse networks with the number of edges approximating the number of nodes, the operation of the Newman (2004a) algorithm is more accurately stated as $O((m+n)n)$ where $n$ is the number of nodes and $m$ is the number of edges in the network being processed. Therefore, the Newman (2004a) algorithm would take approximately 103 times longer to process the complete pilot network as it took to process the sub-network of the pilot network.
A third possibility is that the prediction of identical results set forth in Clauset et al. (2004) is an error. However, analysis of the mathematical reasoning used to justify that conclusion is beyond the scope of this paper.
References


