Citation Analysis of Canadian Case Law

Thom Neale[*]

[*] *Sunlight Foundation, Washington, DC, USA*

Abstract. This study uses simple statistical and functional analysis in conjunction with network analysis algorithms to examine the network of Canadian caselaw using data supplied by the Canadian Legal Information Institute (CanLII). Seeking to explore three basic questions, the study describes the database coverage of CanLII along with that of two commercial vendors and juxtaposes that information with the number of citations to cases decided by courts within each province each year. The study then uses analysis of time-series network rankings for each case to determine 1) the age at which cases in the network typically cease to be important, and 2) what characteristics define those cases that continue to be important despite the passage of time. The analysis reveals that indegree centrality and PageRank scores of caselaw within the network are effective predictors of the frequency with which those cases will be viewed on CanLII's website. Further, statistical and functional analysis of network rankings of each case over time suggest that cases typically cease to be cited in 3 to 15 years, depending on the jurisdiction, with the exception of Supreme Court of Canada decisions, which persist for 50 years. The study concludes that roughly 19% of Canada Supreme Court cases remain important despite the passage of time, whereas in all other jurisdiction, less than 3% of cases continue to be cited regularly over time.

1. Introduction

This study uses simple statistical and functional analysis in conjunction with network analysis algorithms to examine the network of Canadian caselaw using data supplied by the Canadian Legal Information Institute (CanLII). Seeking to explore three basic questions, the study describes the database coverage of CanLII along with that of two commercial vendors and juxtaposes that information with the number of citations to cases decided by courts within each province each year. The study then uses analysis of time-series network rankings for each case to determine 1) the age at which cases in the network typically cease to be important, and 2) what characteristics define those cases that continue to be important despite the passage of time.

The analysis reveals that indegree centrality and PageRank scores of caselaw within the network are effective predictors of the frequency with which those cases will be viewed on CanLII's website. Further, statistical and functional analysis of network rankings of each case over time suggest that cases typically cease to be cited in 3 to 15 years, depending on the jurisdiction, with the exception of Supreme Court of Canada decisions, which persist for 50 years. The study concludes that roughly 19% of Canada Supreme Court cases remain important despite the passage of time, whereas in all other jurisdiction, less than 3% of cases continue to be cited regularly over time.

2. Brief History of Citation Analysis

Citation analysis is an old practice, dating back at least to 1873 when Shepard's Citations first published its index of citation links between court decisions in the United States.[1] Before the rise of computer technology in the latter half of the twentieth century, several individuals proposed new systems of citation indexing that were prescient yet would be largely ignored until the turn of the twenty-first century.

In 1945, Vannevar Bush described a futuristic tool called the Memex, which he imagined would enable information retrieval on an unprecedented scale using microfilm storage in combination with an automated system to

[1] Malmgren, Staffan (2011), *Towards a Theory of Jurisprudential Relevance Ranking. Using Link Analysis on EU Case Law*, Graduate thesis, Stockholm University.

navigate through thousands of storage volumes. As Staffan Malmgren observes, Bush's concept of "trails", or connections between information in different volumes, is strikingly similar to today's hypertext-based information systems.[2]

Ten years later, Eugene Garfield, now recognized as the founder of the field of bibliometrics, found fault with existing subject indexes in use at the time, instead arguing that an "association-of-ideas" or a "thought-index" would better accommodate changes in terminology and the use of differing vocabularies within fields. Though his words may evoke something grander, such as a knowledge graph or a normalized relational database, Garfield ultimately proposed the introduction of a citation network.[3]

Malmgren describes a number of early suggestions that citation indexes could be used to retrieve relevant case law, including Stephen M. Marx's suggestion that the automatic generation of lists of citing cases would be helpful.[4] As Geist reports, in 1971 the scholar Pranas Zunde identified three broad application areas in which citation indexes could be valuable.[5] The first was in quantitative and qualitative evaluation of scientists, publications, and scientific institutions—the now controversial practice of estimating a researcher's prestige based on his or her impact on a citation network.[6] The second was in the modeling of the historical development of science and technology. A modern example of such an effort would be the research by

[2] Malmgren, chapter 3.2.1, citing Bush, Vannevar, *As We May Think*, The Atlantic, July 1945.

[3] Geist, A. (2009), *Using Citation Analysis Techniques For Computer-Assisted Legal Research in Continental Jurisdictions*, Graduate thesis, University of Edinburgh, p. 66. . Available at: http//papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID1397674_code1087080.pdf?abstractid=1397674&mirid=1.

[4] Malmgren chapter 2.4.3, citing Marx, Stephen M. (1979), *Citation Networks in the Law*, Jurimetrics Journal Vol. 10, pp. 121-137.

[5] Geist p. 66.

[6] Id. p. 68.

Cross et al. analyzing the development of stare decisis by analyzing networks of US Supreme Court (US Supreme Court) precedent.[7] The third was in information search and retrieval, which may be its most powerful application with respect to legal information.

Malmgren further describes the influential work of Colin Tapper, who suggested in 1982 that the similarity of two cases could be estimated by comparing vectors of the citations contained in each case. Interestingly, Malmgren reports that Tapper declined to use the standard cosine distance function to calculate the similarity of the cases' citation vectors, instead using a "custom function designed to take into account aspects of citation practices that are particularly distinguishing—for example, citations to very old cases, cases in other jurisdictions (some of the example cases were from US federal courts, which may cite case laws from other states) or citations from higher to lower courts."[8]

In that same vein, in 1995 Howard Tutle observed that traditional retrieval models overlook the context in which individual legal documents occur. As a solution, he suggested that computer-assisted legal research utilize a data structure featuring linked citations, or a network.[9] As Geist explains, Tutle's advice was not followed at the time,[10] yet during the same period the

[7] Cross, Frank B. et al. (2010), *Citations in the U.S. Supreme Court: an Empirical Study of Their Use and Significance*, University Illinois Law Review, No. 2 p. 491. Available at: http://illinoislawreview.org/wp-content/ilr-content/articles/2010/2/Cross.pdf. See generally Fowler, James H. et al. (2008), *The Authority of Supreme Court Precedent*, Social Networks, Vol. 30, No. 1, pp. 16-30. Available at: http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID1008032_code646904.pdf?abstractid=1008032&mirid=1

[8] Malmgren chapter 3.2.4, citing Tapper, Colin (1981), T*he Use of Citation Vectors for Legal Information Retrieval*, Journal of Law & Information Science,Vol. 1, No. 2, pp. 131-161.

[9] Geist p. 58.

[10] Id. at 58, citing Moens, M.-F. (2007), *Summarizing court decisions: Text Summarization*, Information Processing & Management, Vol. 43, No. 6, pp. 1748-1764.

Internet witnessed phenomenal advances in both academia and industry directed at effective search within hyperlinked environments. "In a sense," Geist observes, "citation analysis methods have only been rediscovered and modified for both network analysis and Web search."[11]

3. Application of Network Analysis to Legal Citations

Several scholars have noted in theoretical terms that caselaw citation networks contain valuable information that generally reflects the relevance of precedent.[12] Michael Gerhardzt observed that the extent and nature of a precedent’s network of citations determine the strength of its constraining power on subsequent cases. He argued further that the authority of a precedent depends on the consistency and uniformity with which other authorities have cited it.[13] These kinds of observations often correspond closely to the processes used by network analysis algorithms, which helps strengthen the evidence that such algorithms can be usefully applied to legal citation networks.

A number of studies have done so. In 2005 Thomas A. Smith examined a network of US Supreme Court decisions and observed that the network was scale-free, or exhibited a power-law distribution, as network theory would predict.[14] In 2007, Fowler et al. (2007) tested methods to identify the most legally central decisions of the US Supreme Court at a given point in

[11] Id. p. 66.

[12] See, e.g., Cross et al. p. 523.

[13] Cross et al., quoting Gerhardt, Michael J. (2008), *The Irrepressibility of Precedent,* North Carolina Law Review, Vol. 86, No. 5, pp. 1279-1297. Available at SSRN: http://ssrn.com/abstract=2306700.

[14] See generally Smith, Thomas A. (2005). *The Web of Law,* San Diego Legal Studies Research Paper No. 06-11. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=642863 (accessed 10 December, 2013)

time.[15] Fowler et al. (2008) later studied how the norm of stare decisis had changed over time in the jurisprudence of the US Supreme Court and sought to identify the doctrine's most important related precedents.[16] In 2010, Cross et al. undertook an empirical analysis of the citation practices of the US Supreme Court justices, seeking to assess why the justices cite cases in their opinions, how they differ in doing so, and how those decisions impact the development of the law. In 2012 Malmgren compared the performance of several network analysis algorithms on a citation network of decisions from the European Court of Justice.[17] In 2012, Clark and Lauderdale used network analysis techniques to develop a statistical model of how a line of reasoning develops through a series of related cases.[18] That same year, Marc van Opijnen evaluated the performance of several network analysis measures on an unprecedented network of 5.6 million citations extracted from case law and scholarly writings from the Netherlands.

All major empirical studies have found network analysis to be an effective technique in identifying authoritative precedent. Several studies have found that simple citation analysis measures like degree centrality are effective predictors of the relevance of court decisions.[19] Other studies have concluded that relevance ranking using link analysis algorithms such as PageRank and Hyperlink-Induced Topic Search (HITS) outperformed conventional measures used to define the importance of US Supreme Court

[15] Fowler, James H. et al. (2007), *Network Analysis and the Law: Measuring the Legal Importance of Precedents at U.S. Supreme Court*, Political Analysis, No. 15, p. 325. Available at http://jhfowler.ucsd.edu/network_analysis_and_the_law.pdf .

[16] Fowler et al. (2008), p. 18 and p. 20.

[17] Malmgren: chapter 5.

[18] Clark, Tom S. and Lauderdale, Benjamin E. (2012), *The Genealogy of Law*, Political Analysis, Vol. 20 No. 3 pp. 330-331. Available at http://userwww.service.emory.edu/~tclark7/genealogy.pdf.

[19] See generally, Opijnen.

cases, such as degree centrality[20] and even expert opinion,[21] and that network analysis can be used to predict which cases will be cited more frequently in the future.[22]

4. Overview of Network Analysis Concepts

Network analysis is an application of graph theory in which information is modeled and analyzed as a graph consisting of a set of nodes (or vertices) and the connections between them, called edges (or arcs). An edge is defined as a set of two nodes. Two nodes so connected are adjacent to one another. Graphs can be either directed or undirected. In a directed graph, edges point in a certain direction and are represented visually as arrows between nodes. In an undirected graph, the connections between nodes are simple lines and lack a specific direction. A graph is acyclic if it contains no cycles, or sequences of edges connecting the same node to itself via other nodes.

Networks have been used to study a diverse array of topics in the social and physical sciences, including the nature of contagious disease transmission, the spread of obesity, and the co-sponsorship of bills in the US Congress.[23] But network analysis is best known for its application to the Internet, which

[20] See Geist, p. 50; Chandler, Seth J. (2005), *The Network Structure of Supreme Court Jurisprudence*, Public Law and Legal Theory Series, University of Houston Law Center No. 2005-W-01, p. 15. Available at: http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID742065_code254274.pdf?abstractid=742065&mirid=1. See generally, Lupu, Tonatan et al. (2012), *Precedent in International Courts: A Network Analysis of Case Citations by the European Court of Human Rights*, British Journal of Political Science. Available at http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2015331_code1021034.pdf?abstractid=1643839&mirid=1; Cross et al.; Opijnen.

[21] Malmgren: chapter 5.

[22] See id.; Fowler et al. (2008), p. 18, 20, pp. 21-22; Lupu et al. p.18, 21, pp. 23-24.

[23] Fowler et al. (2007) p. 344.

provides a helpful example for demonstrating many network theory concepts. Perhaps the most important such concept is that of degree distribution, which measures how the edges in the network are distributed among the nodes. If the distribution follows a power law, such that a small number of nodes have a large number of connections and most nodes have few or no connections, the network is a scale-free network. AN example of a scale-free network is airline routes, in which a few large hubs would service the most traffic.[24] If the edges in the network are instead normally distributed, as a bell-curve, the network is a random network.[25] Highway systems provide a familiar example of a random network, in which most nodes generally share a similar number of connections.

5. Legal Citation Networks Are Scale-Free

Although existing scholarship may disagree on which algorithms perform best in analyzing case citation networks, there is one key proposition on which all studies have agreed: case citation networks are scale-free, which is to say that a very small number of cases receive the most citations, and most other cases are cited infrequently or not at all.[26] This characteristic has been observed on networks of cases by the US Supreme Court,[27] the European Court of Human Rights,[28] and the Austrian Supreme Court, among others. This feature of legal citation networks is important because it shows that legal citations are structured similarly to another notorious scale-free network: the Internet. It thus provides some evidence that

[24] Fowler et al. (2008) p. 14-16.
[25] Fowler et al. (2007) p. 344.
[26] Cross et al. p. 523.
[27] Geist p. 60.
[28] Malmgren, chapter 3.

algorithms known to yield valuable rankings on the Internet will also work for caselaw citation networks.[29]

6. Summary of Network Analysis Algorithms

6.1. Centrality Measures

The simplest indicators of importance within a network count the number of connections to each node. This measure is known as degree centrality, and in the case of directed networks it has two variants, in-degree centrality and out-degree centrality, which count the number of in-bound and out-bound connections to each node. Another less simple centrality measure is eigenvector centrality, which assigns relative centrality scores to all nodes in a network in a way to accords greater weight to connections to high scoring nodes. Yet another centrality measure is betweenness centrality, which is equal to the number of shortest paths from all nodes to all others that pass through that node.

Cross et al. relied solely on centrality measures in their study of US Supreme Court case law and judged them to be a reasonable proxy for case importance.[30] Opijnen later concluded that logarithmically scaled variants of degree centrality were reliable predictors of a case's legal authority and outperformed the unscaled centrality measures typically used by other researchers.[31] Yet others have tested the family of centrality measures and identified drawbacks to using them in analyzing caselaw citation networks. For example, Fowler et al. (2007) note a shortcoming of in-degree centrality in that it treats all inbound citations equally. A citation from a

[29] See generally, Smith, Thomas A. See Geist, p. 63; Malmgren: chapters 3, 5; Fowler et al. (2007), pp. 324-326.
[30] Fowler et al. (2008), p. 6.
[31] Lupu et al. p. 18.

landmark case decided by a review court would be treated the same as a citation from an obscure lower court.[32] Cross et al. observe that out-degree centrality could be similarly criticized for failing to account for outbound citations that are unrelated to precedent, such as cases cited in support of routine procedural points. They also raise the question of outbound citations included for illegitimate reasons, perhaps to obfuscate the true rationale of a decision, for example. Fowler et al. (2007) also proffer technical criticism of eigenvector centrality, which may exhibit a downward bias in assessing the importance of recent cases that have not been cited yet.[33]

6.2. Link Analysis Algorithms

For reasons similar to those stated above, researchers studying Internet search algorithms began mining the link structure of hyperlinked documents for stronger indications of authority and relevance. For example, in 2000, Brian Davison demonstrated that Web pages sharing a link tend to be topically related.[34] Other researchers at the time suggested using the network structure of hyperlinks between documents to locate relevant search results, such that an inbound link from another document affected the authority of the linked document in an amount proportionate to its own relative authority.[35]

In 1999, Jon Kleinberg developed the HITS algorithm, a precursor to Google's well known PageRank algorithm. The HITS algorithm calculated a hub score and an authority score for each document. The hub score represented the document's value as a source of links to other authoritative documents. Conversely, the authority score represented the value of the

[32] Geist, p. 63.
[33] Id.
[34] Cross et al. p. 529.
[35] Opijnen, section 5.

document's content in its general topical area. Hub and authority scores are recursively defined, so that a high authority score results when a document is cited by other documents with high hub scores; and a high hub score results when a document cites documents that have high authority scores. PageRank is a related but different algorithm that endeavors to calculate the probability that a "random surfer" will encounter a given page after repeatedly following a random link on each new page while browsing.[36] The developers of PageRank, Sergey Brin and Larry Page, described it as "an objective measure of [a Web page's] citation importance that corresponds well with people's subjective idea of importance."[37]

6.3. Lack of Consensus Among Previous Studies

In the context of legal citation networks, Fowler et al. (2007) described cases with good hub scores as "outwardly relevant," in that they cite other relevant decisions, and cases with good authority scores are "inwardly relevant," in that they are cited by cases that are outwardly relevant. In their influential study, Fowler et al. (2007) analyzed a network of nearly 27,000 US Supreme Court decisions and concluded that the HITS algorithm produced more accurate estimations of relevance than simple citation counts. Both Chandler and Lupu et al. anecdotally confirmed that important precedents in their respective datasets tend to be cited by many outwardly relevant cases.[38]

Yet studies conducted after Fowler et al. (2007) have failed to confirm that HITS is the most effective algorithm for determining caselaw authority. Opijnen found that his logarithmically scaled variant of degree centrality performed better than HITS and PageRank. Malmgren concluded that PageRank and in-degree centrality performed well, and was surprised by

[36] Malmgren, chapter 3.

[37] Fowler et al. (2007), p. 329.

[38] Cross et al. p. 526.

the comparatively low performance of HITS given its good performance in Fowler et al. (2007). No single measure has emerged as clearly superior to the others, whereas in analysis of the Internet, the link analysis algorithms are generally considered to be superior to centrality measures. The next section discusses data modeling challenges that may account for some of the inconsistencies in the performance of the link analysis methods discussed above on caselaw citation networks.

7. Constructing the Network

Although caselaw citation networks resemble the Internet in that both constitute scale-free networks, there remain significant differences between the two. Chief among these is that the network structure of Web pages is explicitly indicated by the hyperlinks contained in each page. The links are unambiguous and easily readable by software. But in caselaw citation networks, the linkages between documents are not always as easy to discern. Consequently, recent studies have failed to identify document traits that completely and accurately model the edge relationships between the nodes in the network.

7.1. SHORTCOMINGS OF PREVIOUS EFFORTS AT DATA MODELING

Most studies have assumed that the network of caselaw citations can be adequately modeled using only full citations as the edges connecting caselaw documents. This assumption is often unstated and is only apparent on review of the studies' explanation of their methodology. Both Fowler et al. (2008) and Chandler used regular expressions to extract citations from their documents, which indicates their citation detection methods were simple and pattern-based, largely ignoring the larger context of the documents in which the citations were located.[39] Only one recent study appears to have gone beyond merely detecting full citations, endeavoring also to identify textual references to cited cases' captions.[40]

The justification these studies have presented on behalf of this assumption have been dismissive of the notion that citations may have differing degrees of strength or credibility. For example, Fowler et al. (2008) argue that every citation represents a latent judgment by the author that the cited resource is in some fashion legally relevant to the issues raised in the citing document. This assertion is no less valid, they argue, even if the citation distinguishes, disapproves, or denies the relevance of the cited case.[41] Malmgren similarly argues that a citation essentially constitutes an endorsement, an assumption which he explains lies at the foundation of the field of bibliometrics. He acknowledges a number of criticisms pointing out limitations of or problems with this foundational assumption, but broadly refers to the success of Google and PageRank as evidence that these problems are not so great as to cast doubt upon it. In the abstract, these arguments are certainly unobjectionable; examples validating these propositions are not hard to imagine. For example, even if an opinion distinguishes a cited case, then, in all likelihood, at least one of the parties has deemed it relevant enough to include in their papers.[42] Nevertheless, these arguments only serve to rebut criticisms that the studies' networks are overinclusive; the more fundamental criticism they fail to address is that the networks studied were incomplete. The discussion of citation extraction methodology in section 7.4 explains the steps I took to address these kinds of issues.

[39] Fowler et al. (2007), p. 330.

[40] Geist, p. 57, citing Davison, Brian D. (2000), *Topical locality in the Web*. In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, Athens, Greece, pp. 272–279. Available at http://www.cse.lehigh.edu/~brian/pubs/2000/sigir/sigir2k.pdf.

[41] Id. p. 53.

[42] Geist, p. 55-57.

Another possible shortcoming of previous studies is that they neglected to consider an important distinction between link analysis algorithms like HITS and PageRank, which are weight-sensitive, and degree centrality measures, which are not. Both PageRank and HITS will factor initial edge weights into their ranking calculations, such that if a certain edge is deemed more or less credible at the outset, any corresponding augmentation to the edge's initial weight value will be reflected in the final rankings of the nodes in the network. Although it may be understandable that previous studies have ignored this factor—inasmuch as it may implicate prohibitively costly measures to identify the credibility of each citation—carefully addressing it may result in a more accurate model of the network and yield more powerful rankings from link analysis algorithms.

7.2. THE ROLE OF EDGE WEIGHTS IN LINK ANALYSIS ALGORITHMS

The simplest way to use edge weights to reflect the structure of a citation network was identified as problematic by Opijnen, but doesn't appear to have been properly addressed in his study. He observed that multiplicity of citations appeared to be "very relevant" in ascertaining a case's importance. Multiplicity arises when one case cites another multiple times, and, as Opijnen notes, most previous studies have taken no steps to account for it.[43] But ignoring edge weights may result in a less accurate representation of the network than if edge weights were considered. For example, if a citing case cites one case eight times in substantive discussion and another case only once in support of a perfunctory procedural point, it is arguably incorrect to treat the two cited cases evenly for ranking purposes. Doing so inadvertently overstates the rank of the case cited once and equally understates the rank of the case cited eight times. Although Opijnen reports

[43] Id., quoting Brin, S., and Page, L. *The Anatomy of a Large-scale Hypertextual Web search Engine*. Computer Networks (and ISDN Systems), Vol. 30, No.1-7, pp. 107–117. Available at http://ilpubs.stanford.edu:8090/361/1/1998-8.pdf.

calculating weighted variants of his degree centrality measures,[44] there is no indication that his calculations using link analysis algorithms used weighted network edges. The same is true of calculations conducted by Fowler et al. (2007) and Malmgren.

A related respect in which initial edge weights may be important in accurately modeling the structure of caselaw networks relates to outliers, another area of concern identified by Opijnen. As he uses the term, an outlier is a case that is cited very frequently, such as a case setting forth boilerplate language on the standard for summary judgment, dismissal of a complaint, exclusion of evidence in criminal case, or dismissal of an untimely appeal. Such cases tend be cited very frequently to document the court's reliance on uncontested principles, and their inclusion tends to be perfunctory and unrelated to the discussion and citations informing the court's substantive analysis. Without accounting for multiplicity, over time these perfunctory citations may receive inflated rankings compared to citations to cases that courts are quoting, discussing at length, citing multiple times, and ultimately relying on in their legal reasoning.

7.3. Further Applications of Edge Weights

How might one translate these indicators of authority and relevance into edge weights suitable as input for weight-sensitive link analysis algorithms? Clark and Lauderdale argue that "opinions engage and discuss the most legally relevant precedent the most."[45] Although this view has been and called "facile" by Cross et al., who disagree that judges mechanistically base their decisions on the most objectively relevant precedents,[46] it may prove valuable with a small modification. Instead, we might choose to be agnostic about whether judges discuss objectively relevant cases, and

[44] Chandler p. 15; Lupu et al. pp. 20-21.
[45] Fowler et al. (2008), p. 18; Chandler pp. 3-7.
[46] Clark and Lauderdale, p. 333.

simply observe that opinions engaged and discussed at the greatest length ought to be the most highly ranked.

This proposition is, so far as my research has revealed, entirely unexplored. Only one study appears to have attempted to distinguish "strong" citations from weaker citations.[47] In that study, David Walsh defined a citation as strong if (1) it directly quoted a cited case and the quotation's length exceeded a single word or phrase, (2) the discussion of the case exceeded a single sentence in length, or (3) the citing court explicitly articulated reliance on the cited decision. Walsh calculated the degree centrality across a network consisting only of strong citations and compared it to the degree centrality across a corresponding network of all citations. Walsh found no significant advantage to using only strong citations in his calculations.[48] But his dataset was very small—a mere 157 cases were examined—and he only used his observations concerning the strength of citations to eliminate supposedly uninformative edges from his network, not to apply weights to the citation edges in his network. At minimum, Walsh's findings ought not to deter efforts to analyze caselaw citation networks with weighted link analysis algorithms.

When citation network edges are accorded weights proportional to the extent of the discussion they receive in citing documents, we might expect to see two resulting benefits. First, we could simultaneously control for Opijnen's outliers, which are unlikely to be discussed at length because, by hypothesis, those citations are usually perfunctory references and receive minimal discussion. And second, we could account for the influence of multiplicity by literally multiplying the weight of the cited document's network edge by some factor proportional to the number times it was cited. Furthermore, by identifying other contextual indicators of extended

[47] Fowler et al. (2008) p. 18; see also Cross et al. p. 494.
[48] Malmgren, chapter 3.

discussion, such as multiple citations, block quotations, inline quotations, short citations, and textual references to case titles and title fragments, it might be possible to derive an even richer set of features with which to model the actual relevance structure of a caselaw citation network.

7.4. Methodology

In this study, I endeavor to identify these objective indicators of extended discussion within a corpus of 594,540 Canadian court opinions and analyze the resulting citation network with the aim of ranking the cases in order of relevance to future legal researchers.

7.4.1. *Citation Extraction*

I used the following methodology to extract citations from the full text of court opinions in a way that considers the overall document context in which they occur. First block quotations were identified by opening each decision in a headless Firefox browser and querying each paragraph for the computed width of its left-hand margin. Indented text was assumed to constitute a block quotation if the preceding flush paragraph ended in a colon or em dash and a citation was located in the flush paragraph text immediately preceding or following the indented text. Next, a context-free grammar was used to extract inline quotations using a similar heuristic. If the quotation was unambiguously preceded or followed by a case citation, the two were presumptively related. Even if these methods weren't perfectly reliable at determining whether a particular quotation originated from an adjacent citation, the proximity of the two may nevertheless indicate that the cited case is relevant to the citing document's analysis. For example, the quote could simply be language from a party's brief, dialog from a transcript, a characterization of a party's argument, or (even better) a quoted statute or regulation. In those circumstances, an adjacent case citation may bespeak a refutation of the quoted proposition or a validation of it. Just as a disapproving or distinguishing citation can still reflect informative judgments about precedent, so might citations positioned adjacent to quotations, whatever the actual source of the quotations may be.

Next, full citations were extracted using a context-free grammar calibrated to identify 112 different strings used in Canadian caselaw citations and the surrounding volume, page, year, and slip opinion numbers that typically accompany them. The parser was capable of identifying pinpoint page citations, page range citations, footnote citations, subsequent history citations, and parallel citations. A parse tree was constructed for each case, reducing it to nodes representing paragraphs, quotations, and sources accompanied by one or more citations and their component parts.

After the parse trees were created, a software technique known as the visitor pattern was used to traverse the parse trees and scan backwards from full citations to isolate the full title of each case. Then fragments of the case title were precomputed and a second visitor was used to identify short citations and textual references to case titles in other paragraph text nodes. The short citations and case titles were then resolved back to full citation nodes, if possible. The rationale for identifying short citations and textual references to titles and title fragments is simple: in the same way that people use first names and nicknames to refer to friends and others with whom they are more familiar, authors of legal text use short citations and titles to refer to sources that are more "familiar" to their analysis, i.e., sources they refer to more frequently, suggesting those sources are more factually or doctrinally related to the instant case than other sources cited less frequently in the document.

The final result of this process is a parsed syntax tree representing the entire body of the opinion. The root node of the tree represents the full text of the decision and has one or more child nodes representing paragraphs within the document. Each paragraph has one or more child nodes representing chunks of text content, full citations, short citations, or quotations. Each of those in turn has any of a number of defined child nodes. The full citation nodes have a "Title" child node with the title of the decision and a "Citations" child node with one or more children representing the various parallel forms of citation that refer to the cited source. As an illustration, consider the following paragraph:

> There is some case law suggesting (without much discussion) that a purchaser cannot maintain a caveat unless it can be shown that specific performance is available. Where there is no binding contract, such that the purchaser is unable to get any remedy, clearly a caveat cannot be maintained: Oxford Development Group Inc. v. Midland Development Ltd., [1993] A.J. No. 47 (C.A.).

Below is the same paragraph reduced to a parse tree (the full origin paragraph and parse tree are shown the Appendix):

```
-Start([])
  -Node([])
    -Content([(0, Token.Content, u'There is
some case law suggesting (without much discussion) that a purchaser
cannot maintain a caveat unless it can be shown that specific
performance is available. Where there is no binding contract, such
that the purchaser is unable to get any remedy, clearly a caveat
cannot be maintained: ')])
    -Source([])
      -Title([(297, Token.Title, u'Oxford Development Group Inc. v.
Midland Development Ltd.')])
      -Citations([])
        -Citation([])
          -SlipYear([(356, Token.SlipYear, u'[1993]')])
          -Reporter([(363, Token.Reporter, u'A.J. No.')])
          -SlipNumber([(372, Token.SlipNumber, u'47')])
          -Jurisdiction([(375, Token.ParenAbbrev, u'(C.A.)')])
    -Content([(381, Token.Content, u'; ')])
```

7.4.2. *Shortcomings of Methodology*

This citation extraction methodology described above is quite powerful, but several potential shortcomings deserve mention.

7.4.2.1. *Challenges of Title Extraction.* The first is that the title extraction method of scanning backwards from the bound volume citation requires much fine tuning and special casing in order to achieve good results across a large corpus. Without explicit programming the parser to include or exclude certain phrases from the title—particularly, phrases dense with acronyms—the parser will either stop prematurely, yielding an incorrectly truncated title, or continue lexing too far backwards, incorrectly including non-title prose preceding the title. I spent many hours tuning this feature and testing it, and although it worked well, I ultimately opted not to rely on the extracted titles because the number of incorrectly parsed titles was higher than I had hoped.

Accurate case titles are important to the process of detecting and merging two distinct citations that refer to the same case, and in the absence of canonical data on parallel citations, a full-featured citation resolution system would have to make use of titles for this purpose. Consequently, the network had some amount of duplication in it, meaning that some sources were represented by multiple distinct source nodes in the network, and their network ranking scores where artificially distributed among those nodes rather than consolidated into a single accurate ranking. The regression analysis below suggests that the network modeled the database collections reasonably well in spite of this duplication, but there is room for significant improvement in the citation resolution methods used.

7.4.2.2. *English versus French Cases.* A second likely shortcoming of this methodology is that in fine-tuning the citation extraction code, I focused mainly on decisions written in English, so it is possible that a certain percentage of citations weren't detected in decisions written in French. I tested the code on many Quebec cases, but as the database coverage charts below demonstrate, three of the largest databases under examination exclusively held cases from Quebec. Though not all cases in those database were in French, a significant number were, and it's possible that the citation extraction routines were less effective on French decisions, and that any such flaws were amplified by the sheer size of the Quebec-specific caselaw databases. The chart comparing Quebec's indegree density versus database coverage (explained in section 10.3 below) shows very few citations detected prior to 1990, which may provide indirect evidence of this shortcoming.

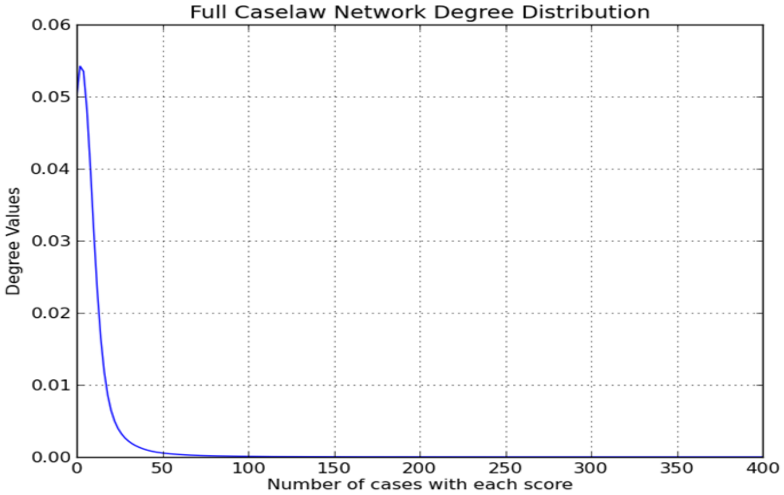
8. Network Analysis

8.1. Size and Scope

The full case law network consisted of 1,900,916 citations distributed among 566,992 nodes, a concept that corresponds roughly to individual source documents, though imperfectly. Roughly 40% of these nodes resolve directly to cases in CanLII's database. The remaining 60% represent citations to cases beyond the current scope of CanLII's collections or were citations to unofficial reports, like the Criminal Reports, that weren't always possible to resolve back to CanLII documents.

8.2. Degree Distribution

Overall, the distribution of the citation "edges" among the source nodes in the network adheres to the power law distribution predicted by the literature, which is to say that a small number of cases receive a large number of citations, and a large number of cases receive few citations or none at all.



8.3. Selection of Graph Analysis Algorithms

To determine which graph analysis algorithms would yield the most reliable information, I computed a rank for each node in the network using indegree centrality, outdegree centrality, eigenvector centrality, PageRank, and Hyperlink-Induced Topic Search (HITS). Using linear regression, these ranks were compared to the number of page views each matched case received on CanLII's website during the year 2012.

The decision to use page views as the external benchmark in the regression analysis deserves some explanation. Page views arguably may not reflect legal importance or relevance as much as they reflect topical frequency. A case may have mild legal relevance within the whole corpus of cases, but will be viewed frequently on the website if its subject matter is commonly shared with legal issues that users of the website need to research. This measurement is likely to be biased in favor of the same routine cases that benefit from the bias inherent in indegree centrality.

The decision to use page views was motivated by two main factors. First, better external data points concerning legal relevance were simply unavailable. In the course of this research, I explored many options for obtaining authoritative sources on Canadian law, such as treatises and legal newspapers and magazines, but all proposed uses were precluded by the publishers' restrictive licensing and use provisions. And second, a more difficult question is whether legal relevance is truly distinguishable from topical frequency. If users are measurably more interested in certain cases, what is the value of a countervailing notion of legal relevance across an entire dataset? This is a question I felt ill-equipped to address in the time available, but it may be an important one, and certainly merits further research. In any event, until then we have little choice but to rely on page views.

Algorithm	**r-value**	**p-value**
Indegree	0.39	0.0
PageRank	0.35	0.0
Outdegree	0.23	2.56
HITS hub	0.22	9.90
Eigenvector	0.09	2.92
HITS auth	0.03	4.84

With the exception of eigenvector centrality and HITS-authority, all algorithms yielded statistically significant correlations with page views. The correlation coefficients for indegree and PageRank are particular striking. Although I believed PageRank and HITS might outperform indegree centrality at the outset of this project, I decided to use indegree centrality scores for the remainder of the computations in my report after reviewing these figures. The correlation is far too strong to ignore, and even if indegree centrality tends to inflate the rankings of cases that stand for minor or routine points of law, the regression analysis shows that users are highly interested in those cases.

8.4. Findings Regarding Edge Weights

The network algorithms in the table above exclude any calculations using weighted network edges due to what appeared to be an error in the calculations using the weighted network. The PageRank scores were identical regardless of how the network edges were weighted, which may suggest a flaw in my implementation of PageRank. This is still an interesting and potentially valuable avenue of research, but those results were unusable, so I excluded them from the remainder of the analysis.

9. Exploring Database Coverage

To better understand the scope of the caselaw collections of CanLII, Westlaw, and LEXIS, and how each relates to citation trends apparent within the caselaw network, the next sections describe the database coverage of each, then plots each against a histogram showing the number of cases cited in each jurisdiction, broken down by publication year.

9.1. Current Database Coverage

9.1.1. *Westlaw*

Westlaw endeavors to provide "coverage of unreported court decisions from 1986 forward and reported court decisions from 1977, as well as decisions published in Carswell Law Reports from their inception."[49] Its collection also includes decisions predating 1977 from "key courts and law report series." These mostly include archival decisions from discontinued reporters, such the Alberta Law Reports between 1908 and 1933. Westlaw's collections include the full archives of several unofficial reports, such as the Reports of Family Law (1824 to present), the Western Weekly Reports (1911 to present), and the Criminal Reports (1946 to present).

[49] Opijnen, section 1.

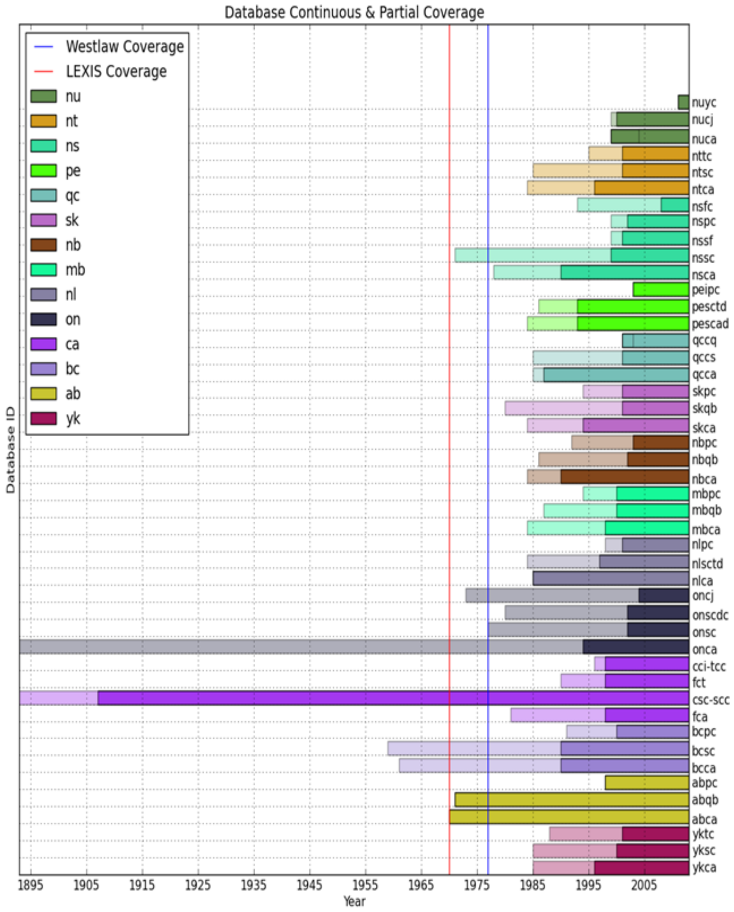
9.1.2. *LEXIS*

LEXIS apparently does not publish a similar document detailing the scope of its case law collections. A LEXIS representative told me that their collections include all published cases from 1970 onward, plus many unpublished cases. In addition, the representative told me that any case cited in the previous set of cases that is dated before 1970 is also present in the LEXIS databases. In the network I constructed from CanLII's collections, roughly 60 percent of all cases were never subsequently cited, so a reasonable speculation might be that LEXIS has roughly 40% coverage of cases published before 1970.

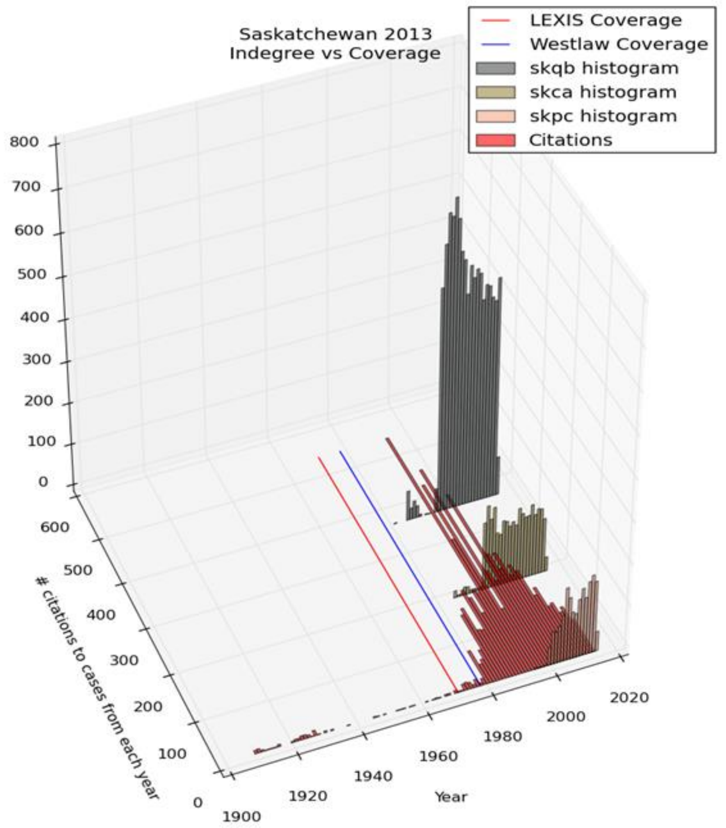
9.1.3. *CanLII*

CanLII publishes detailed information on the scope of its collections.[50] The stacked bar chart below depicts CanLII's continuous coverage of each database as solid colored bars. The right-hand edge of the chart represents the year 2013, with colored bars extending leftwards, back in time. Extending beyond the ends of the solid colored bars are slightly transparent bars that indicate the scope of partial coverage, which was calculated using CanLII's API. Also visible in the chart are two vertical lines, one blue and one red, representing the general boundaries of the vendors' continuous coverage of published decisions.

[50] This point is puzzling, as degree centrality measures are typically insensitive to edge weights.


Database Continuous & Partial Coverage
Westlaw Coverage
LEXIS Coverage
nu
nt
ns
pe
qc
sk
nb
mb
nl
on
ca
bc
ab
yk
Database ID
nuyc
nucj
nuca
nttc
ntsc
ntca
nsfc
nspc
nssf
nssc
nsca
peipc
pesctd
pescad
qccq
qccs
qcca
skpc
skqb
skca
nbpc
nbqb
nbca
mbpc
mbqb
mbca
nlpc
nlsctd
nlca
oncj
onscdc
onsc
onca
cci-tcc
fct
csc-scc
fca
bcpc
bcsc
bcca
abpc
abqb
abca
yktc
yksc
ykca
1895
1905
1915
1925
1935
1945
1955
1965
1975
1985
1995
2005
Year

Using these total coverage figures as a baseline, the next section examines the efficacy of these three differing degrees of coverage in light of the practices and trends revealed by the citation network.

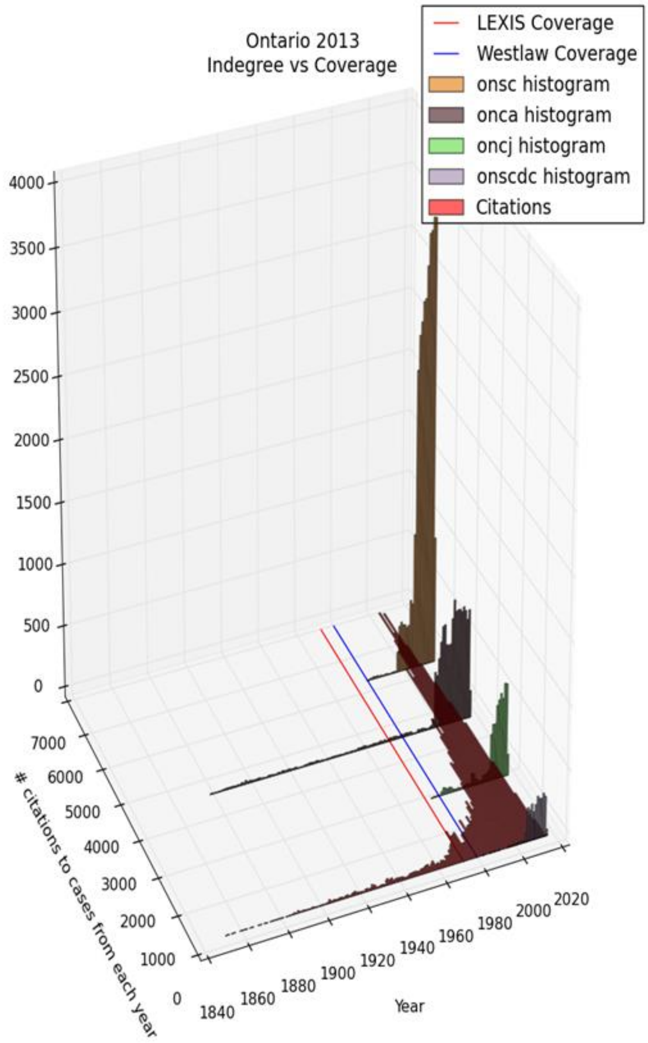


9.2. Coverage vs Indegree Density per Year

The charts in this section are histograms representing database coverage by year. Each vertical bar series represents the number of cases (shown on the vertical *z*-axis) in CanLII's database in the given jurisdiction during each year indicated along the *y*-axis. Another histogram lies flat across the bottom of the chart. That chart plots the sum of all 2012 indegree rankings of citations to cases decided each year, also represented by the same *y* axis. The *x*-axis, to the right, indicates the number of citations to decisions published in each year.

There are two main caveats to keep in mind when interpreting this data. First, the number of citations corresponding to the *x*-axis is approximate. In deriving it, I avoided counting CanLII's custom citation styles (like 2005CanLII23456) in an effort prevent the numbers from being skewed upward during the period of CanLII's continuous coverage due to double counting. I focused on counting citations to bound volumes, with the goal of producing a more realistic distribution of cited cases by year. The data exhibits the features we expect—citations to recent cases are more numerous, and as the year approaches 2013, a sharp drop-off occurs, since very recently decided cases have not been cited at all yet.

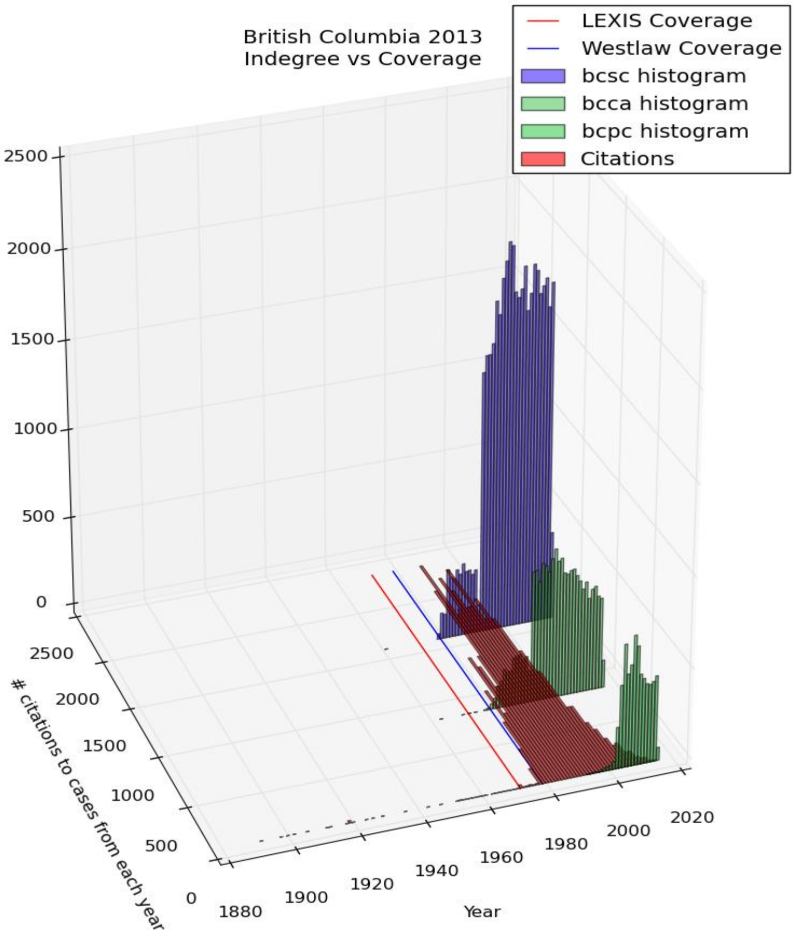
The second caveat is that the density of detected citations per year is determined by at least two unrelated factors. The first is the extent of coverage in the collection from which the citations were extracted. The second is the extent to which courts cite documents published in each year. Consider the plot below for Ontario.

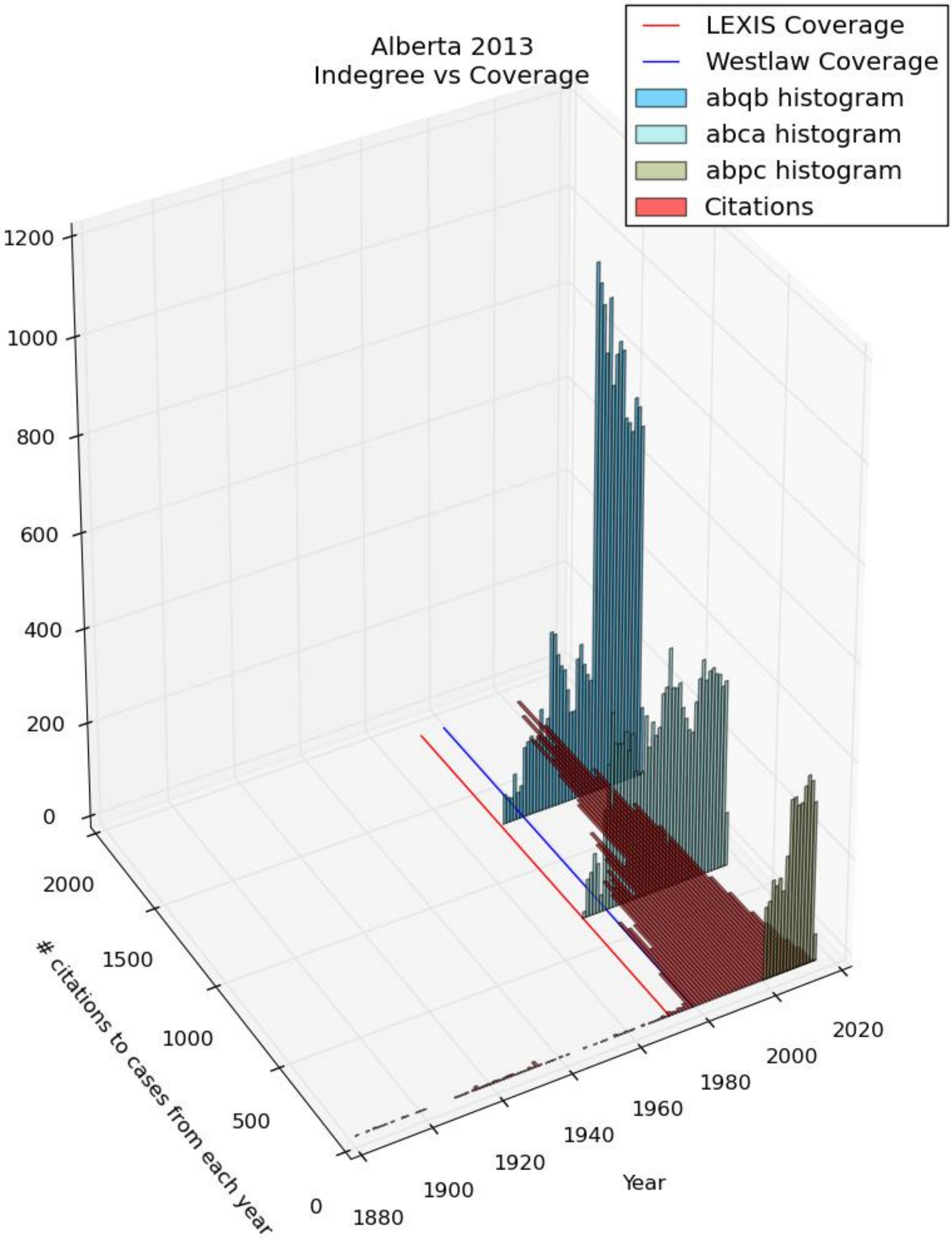

Ontario 2013
Indegree vs Coverage
LEXIS Coverage
Westlaw Coverage
onsc histogram
onca histogram
oncj histogram
onscdc histogram
Citations
4000
3500
3000
2500
2000
1500
1000
500
0
7000
6000
5000
4000
3000
2000
1000
0
# citations to cases from each year
1840
1860
1880
1900
1920
1940
1960
1980
2000
2020
Year

The chart reveals a distinctive long tail of cited cases that extends back to the late nineteenth century. But before concluding that Ontario courts cite older law more frequently than other provinces, in most of which no such tail is present, we must consider that Ontario's database coverage goes back much farther in time than the other provinces. Specifically, the chart shows that the purple histogram representing the ONCA database's coverage has a similar tail that runs parallel to the indegree histogram's tail. The two are probably related, in that the citations detected during that period probably came from the full text of those older cases. Yet if we return to the respective Saskatchewan chart at the beginning of this section, there we also notice a long tail indicating citations to older cases; but in contrast to Ontario, we see no corresponding historical coverage to explain it. In Saskatchewan, therefore, it seems more likely that the long tail of citations to older cases is due to modern courts citing historical case law.

9.3. Historical Coverage May be More Important in Smaller Jurisdictions

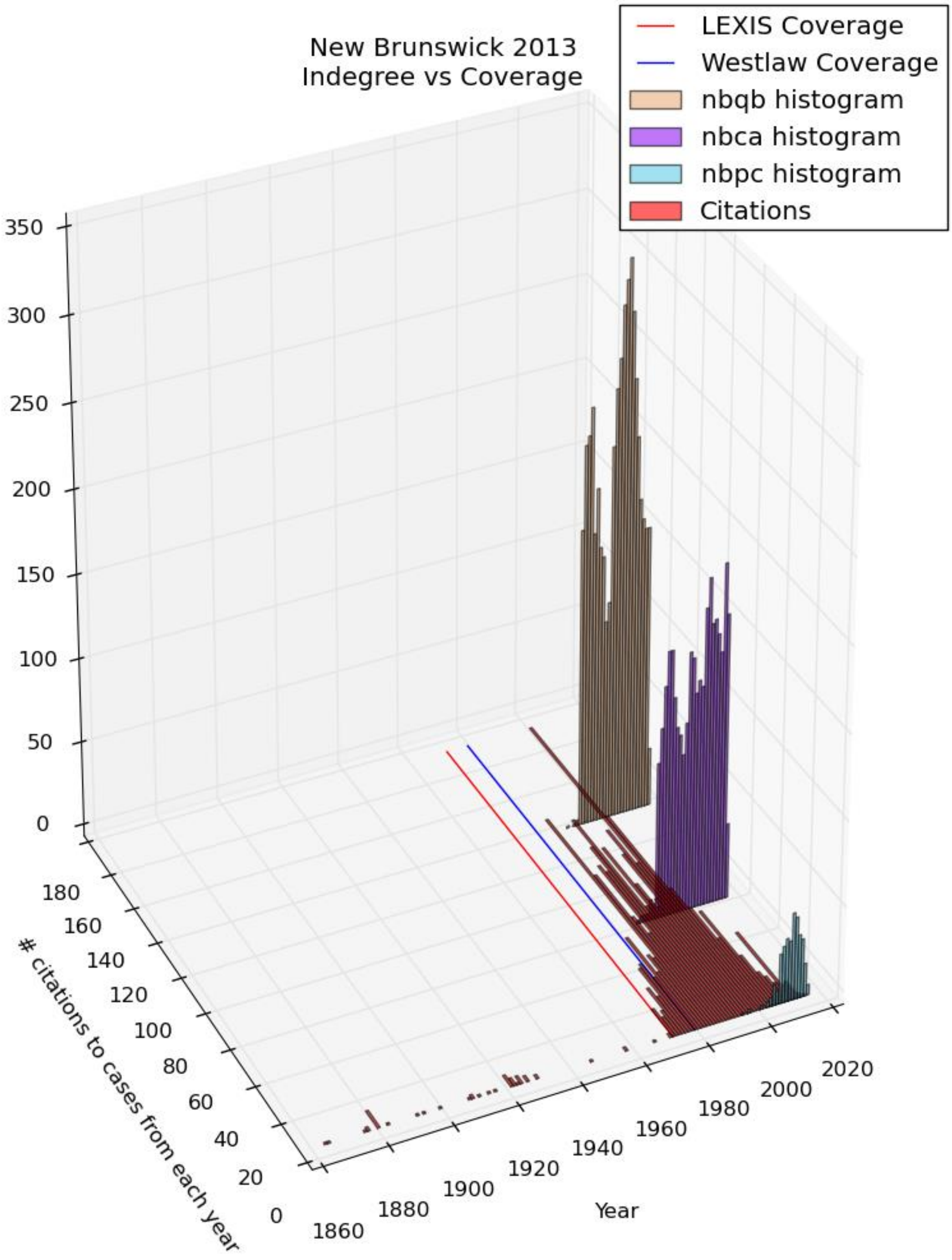
Keeping in mind the caveats explained above, the data seems to show that in the most populous jurisdictions—Ontario, Quebec, British Columbia, and Alberta—courts cite older precedent less frequently than they do in smaller jurisdictions. As examples, consider the charts below showing citation density per year versus database coverage for British Columbia and Alberta.

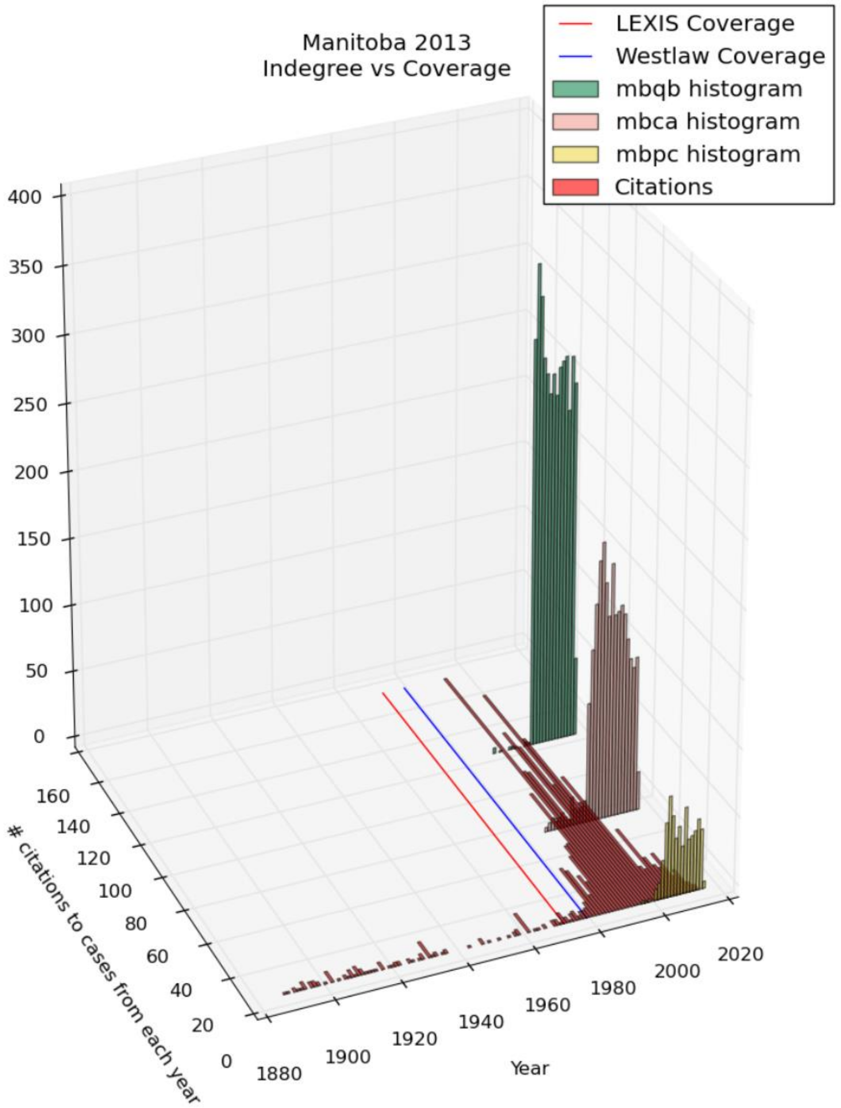

British Columbia 2013
Indegree vs Coverage
LEXIS Coverage
Westlaw Coverage
bcsc histogram
bcca histogram
bcpc histogram
Citations
2500
2000
1500
1000
500
0
# citations to cases from each year
Year
1880
1900
1920
1940
1960
1980
2000
2020


Alberta 2013
Indegree vs Coverage
LEXIS Coverage
Westlaw Coverage
abqb histogram
abca histogram
abpc histogram
Citations
1200
1000
800
600
400
200
0
2000
1500
1000
500
0
# citations to cases from each year
1880
1900
1920
1940
1960
1980
2000
2020
Year

In each, the density of detected citations drops precipitously right around the 1970-1977 period that marks the beginning dates of LEXIS and Westlaw's continuous coverage, respectively. This partially reflects the lower number of cases in CanLII's collections toward those earlier time periods, providing less full text for citation extraction. But it also may demonstrate that courts in those jurisdictions tend to cite older precedent infrequently, which makes sense, considering that lawyers generally favor citing recent precedent when available. Perhaps these larger jurisdictions have an abundance of relevant precedent to choose from, and therefore preferentially cite that newer body of precedent in accord with familiar legal citation values.

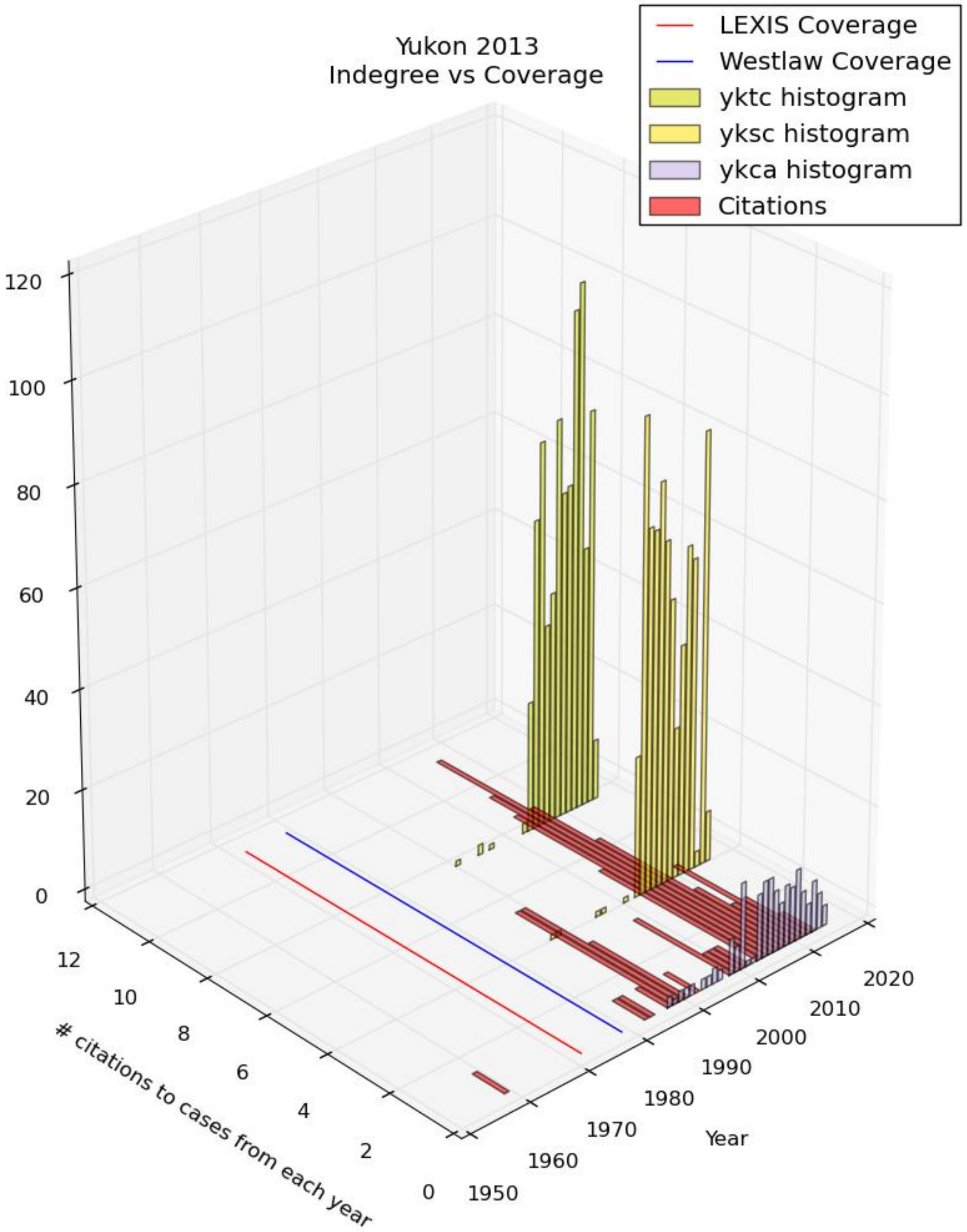
The next four jurisdictions in order of population size are Manitoba, Saskatchewan, New Brunswick, and Nova Scotia. The largest of these, Manitoba, is nearly one third the size of the smallest jurisdiction in the previous group, Alberta. The charts for these jurisdictions exhibit a subtle difference. For example, consider the charts for New Brunswick and Manitoba shown below.

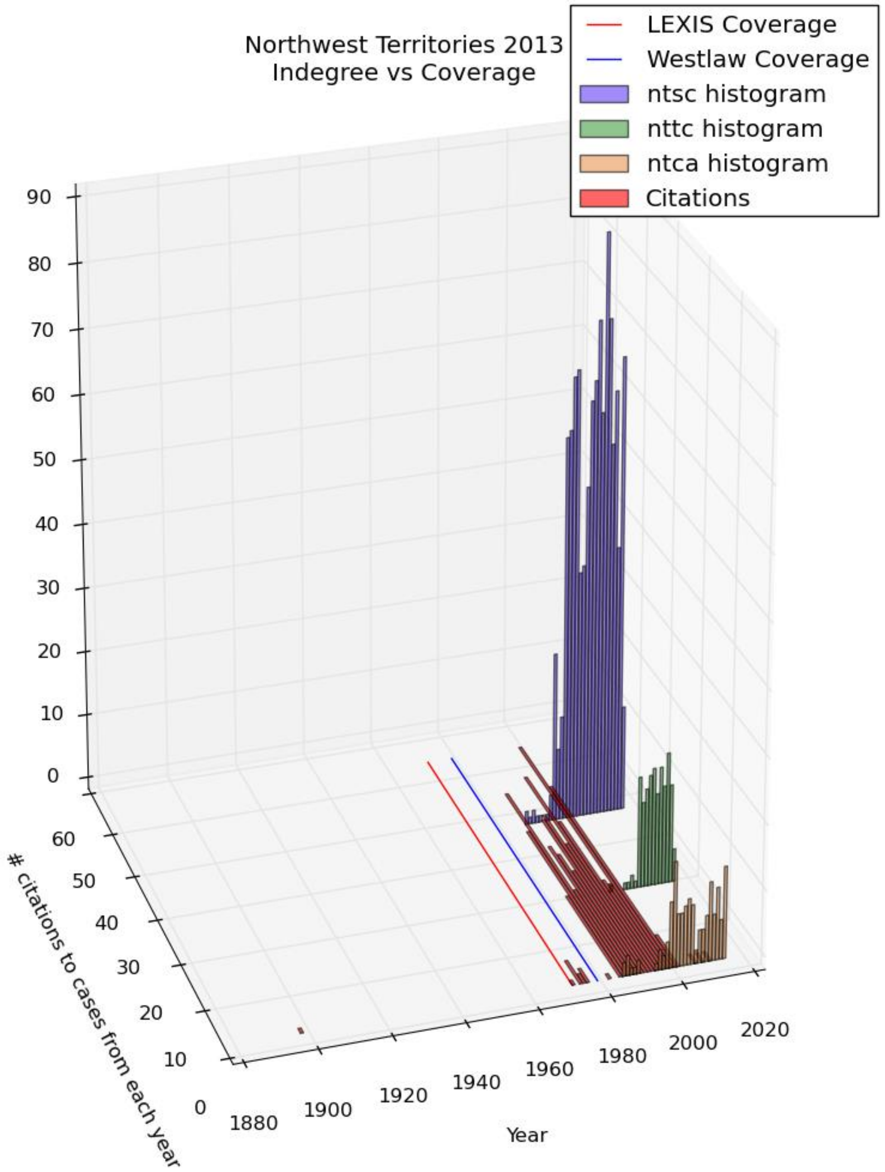

New Brunswick 2013
Indegree vs Coverage
LEXIS Coverage
Westlaw Coverage
nbqb histogram
nbca histogram
nbpc histogram
Citations
350
300
250
200
150
100
50
0
180
160
140
120
100
80
60
40
20
0
# citations to cases from each year
1860
1880
1900
1920
1940
1960
1980
2000
2020
Year

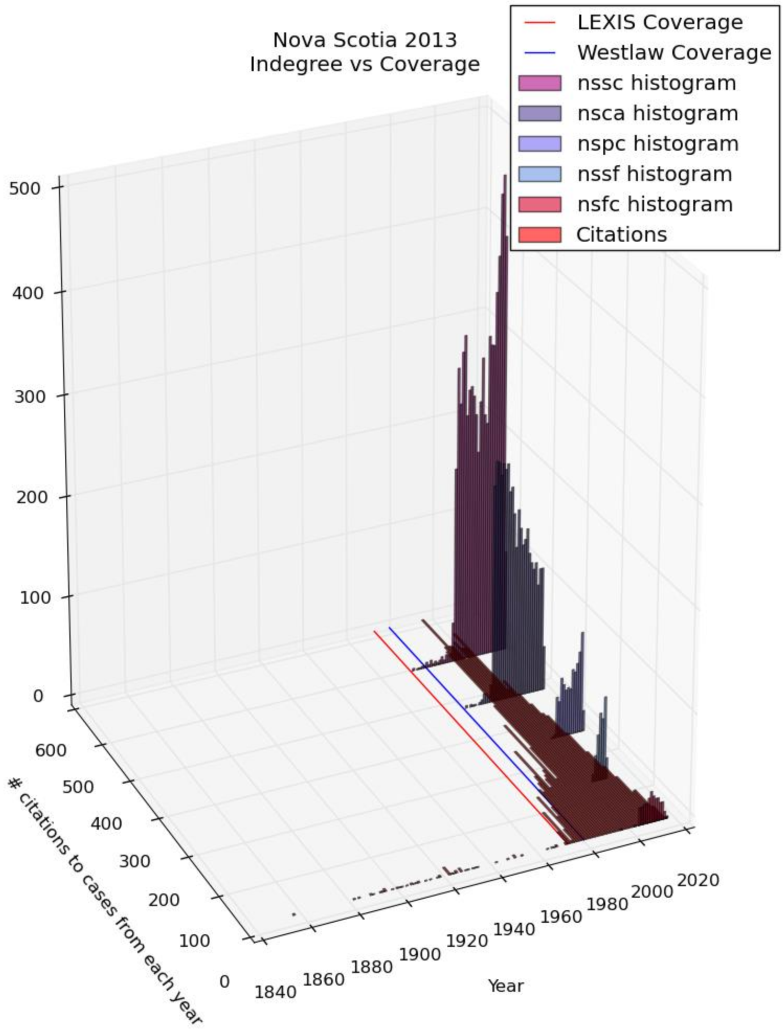

Manitoba 2013
Indegree vs Coverage
LEXIS Coverage
Westlaw Coverage
mbqb histogram
mbca histogram
mbpc histogram
Citations
0
50
100
150
200
250
300
350
400
# citations to cases from each year
0
20
40
60
80
100
120
140
160
Year
1880
1900
1920
1940
1960
1980
2000
2020

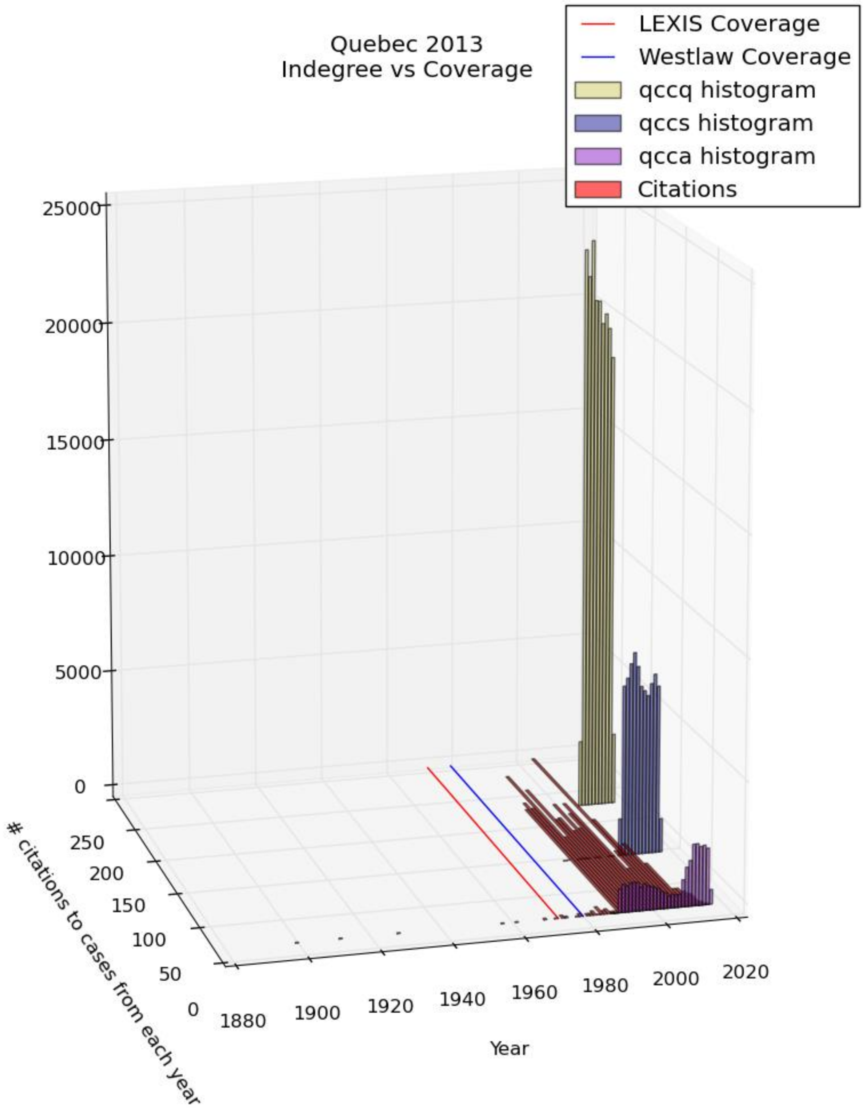
In these jurisdictions, the density of detected citations appears to be skewed noticeably farther left than was evident in the larger jurisdictions. If we were to fit a regression line to this data and examine the line's slope, it may reveal that the larger jurisdictions trend more strongly toward citation of recent cases while the smaller jurisdictions continue to derive utility from their older cases for years. Continuing along the same line of reasoning above, perhaps these less populous jurisdictions simply have less comprehensive case law, such that finding a case that is on point for a particular issue is much more difficult. This would at least explain why judges appear to reach farther back in time to cite older cases; even though recent precedent is more persuasive, sometimes it simply isn't available.

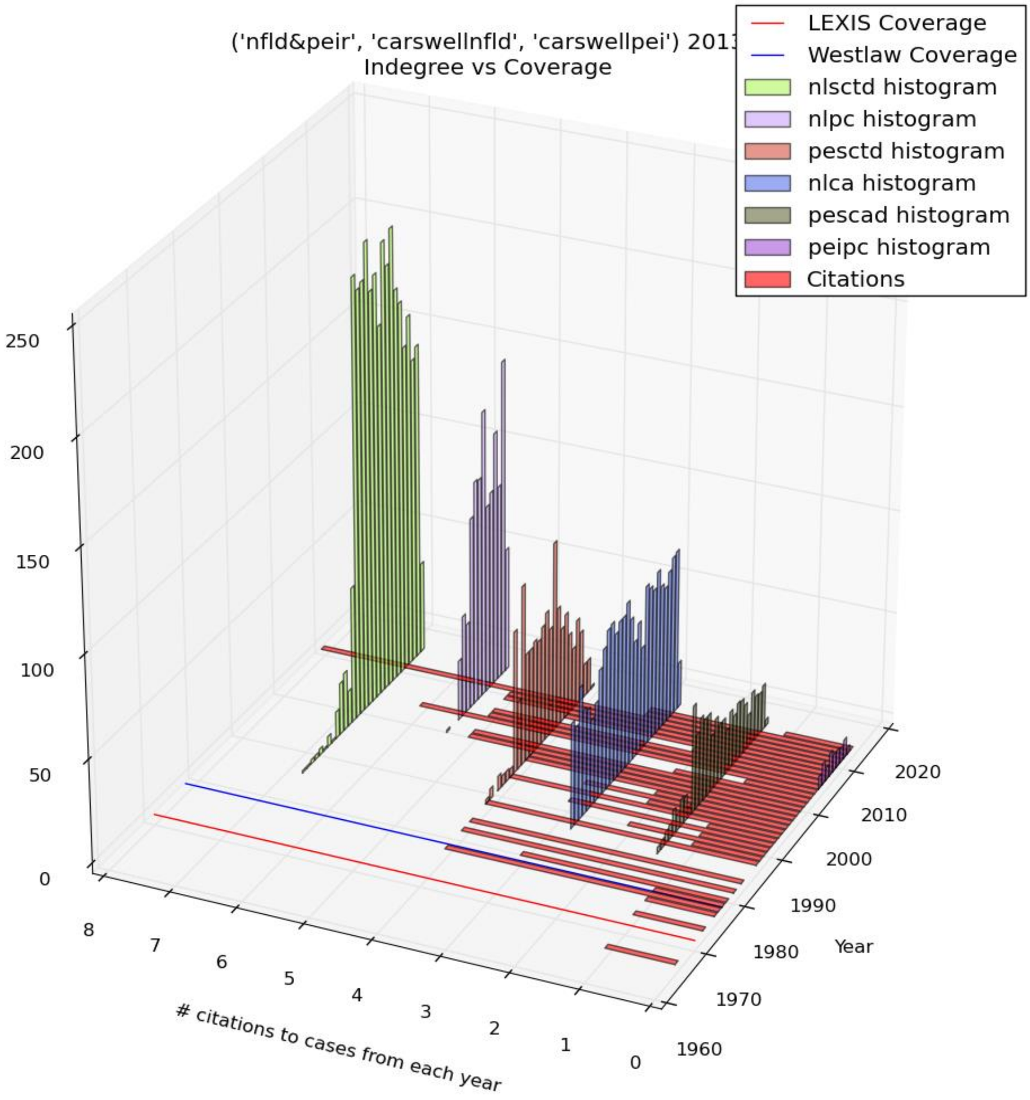
If this is true of these four mid-sized jurisdictions, then we might also extrapolate that older precedent is even more important in the smallest jurisdictions in Canada. Charts for the remaining provinces are shown below, with the exceptions of Nunavut, which had insufficient data to generate a chart. Note also that Newfoundland and Prince Edward Island are combined into a single chart.


Yukon 2013
Indegree vs Coverage
LEXIS Coverage
Westlaw Coverage
yktc histogram
yksc histogram
ykca histogram
Citations
0
20
40
60
80
100
120
12
10
8
6
4
2
0
# citations to cases from each year
1950
1960
1970
1980
1990
2000
2010
2020
Year

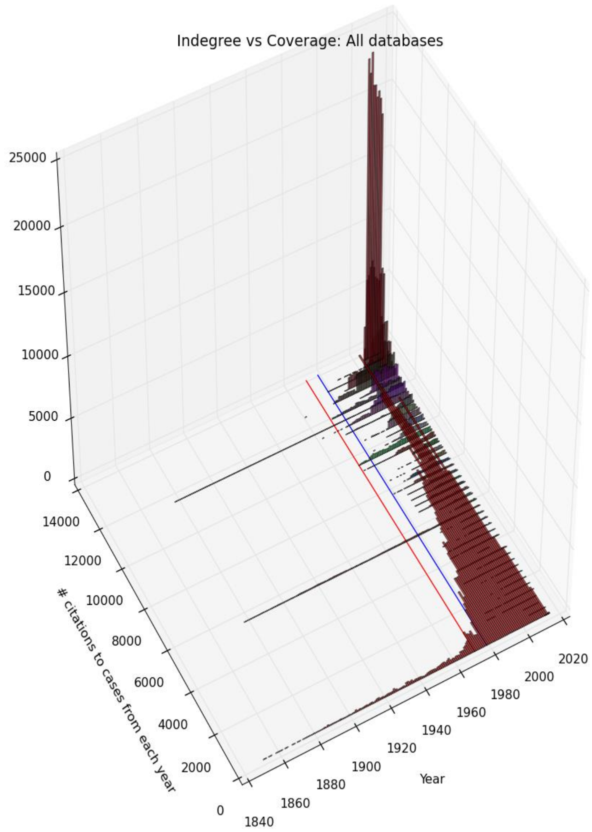

Northwest Territories 2013
Indegree vs Coverage
LEXIS Coverage
Westlaw Coverage
ntsc histogram
nttc histogram
ntca histogram
Citations
0
10
20
30
40
50
60
70
80
90
# citations to cases from each year
0
10
20
30
40
50
60
Year
1880
1900
1920
1940
1960
1980
2000
2020


Nova Scotia 2013
Indegree vs Coverage
LEXIS Coverage
Westlaw Coverage
nssc histogram
nsca histogram
nspc histogram
nssf histogram
nsfc histogram
Citations
500
400
300
200
100
0
600
500
400
300
200
100
0
# citations to cases from each year
1840
1860
1880
1900
1920
1940
1960
1980
2000
2020
Year


Quebec 2013
Indegree vs Coverage
LEXIS Coverage
Westlaw Coverage
qccq histogram
qccs histogram
qcca histogram
Citations
25000
20000
15000
10000
5000
0
250
200
150
100
50
0
# citations to cases from each year
1880
1900
1920
1940
1960
1980
2000
2020
Year


('nfld&peir', 'carswellnfld', 'carswellpei') 2013
Indegree vs Coverage
LEXIS Coverage
Westlaw Coverage
nlsctd histogram
nlpc histogram
pesctd histogram
nlca histogram
pescad histogram
peipc histogram
Citations
250
200
150
100
50
0
8
7
6
5
4
3
2
1
0
# citations to cases from each year
2020
2010
2000
1990
1980
1970
1960
Year

Finally, below is a chart depicting these same metrics in the aggregate across all jurisdictions. Generally, the chart appears to suggest that the majority of citations in the current CanLII collections are to cases published in 1970 or later, suggesting that the industry practices concerning database coverage followed by Westlaw and LEXIS present sound general guidelines that CanLII could follow. Once again, however, it is important to note that this analysis only used a partial dataset, and without a broader corpus of historical cases to analyze, there is no way to be certain that the higher density of citations to recent cases does not simply reflect the boundaries of coverage in CanLII's collections. Similarly, another important question in confirming whether courts in smaller jurisdictions cite older cases more frequently would be to examine the extent to which those courts cite cases from other jurisdictions compared to their own cases. If relevant local precedent is truly rarer in smaller jurisdictions, we might expect to see courts in those jurisdictions cite cases from other jurisdictions with a greater than average frequency.

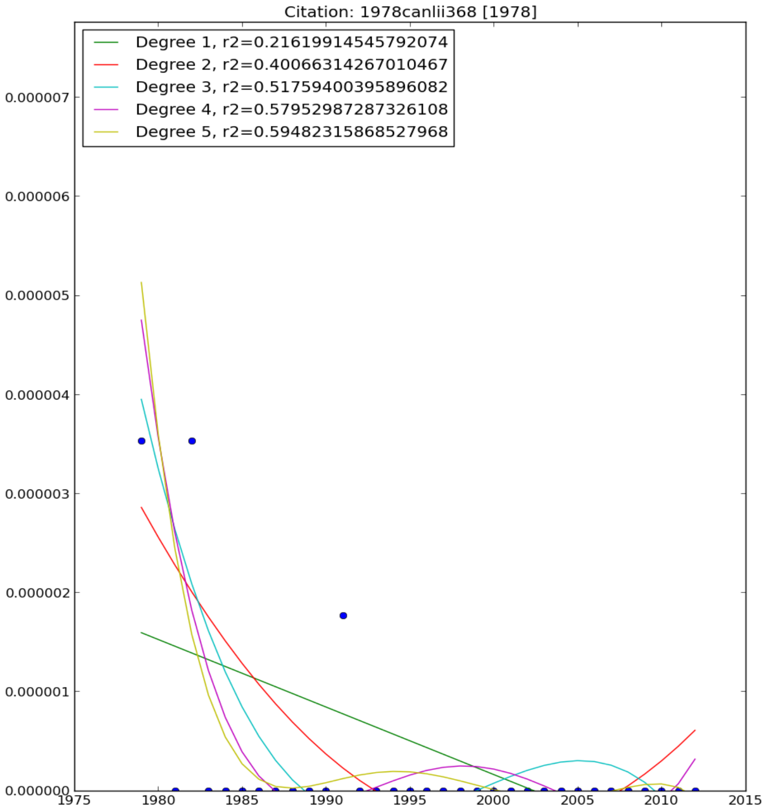


*Citation density by year versus database coverage. The legend is omitted because it would be too large to display***.**

10. Calculating the Age at which Cases Cease to Be Important

10.1. Methodology

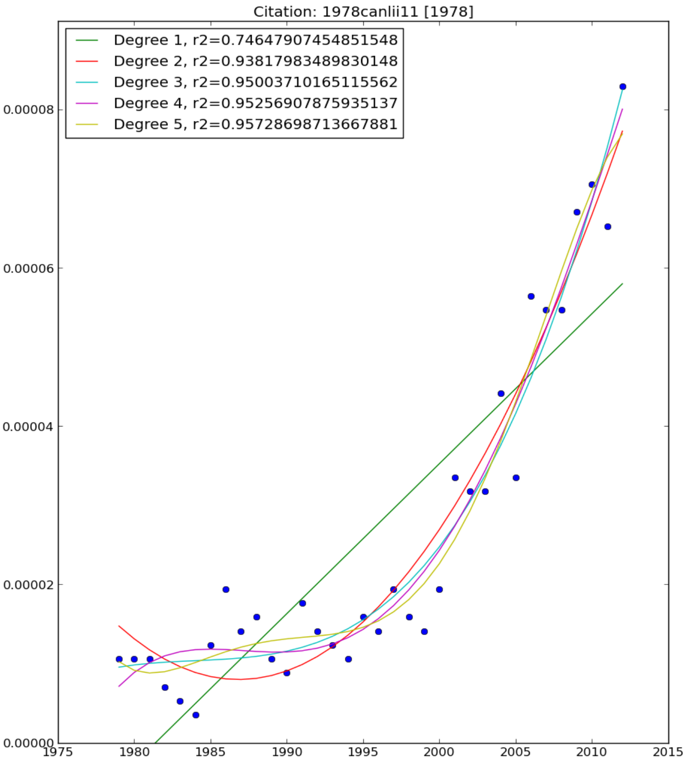
Determining the age at which cases cease to be important requires 1) some way to quantify the importance of each case at a given point in time, and 2) some way to measure the overall trend of changes in the case's importance over time. Using the year-by-year indegree centrality scores for each case computed during the network analysis phase, I used a linear curve-fitting algorithm to obtain a regression line, then queried the line for its slope to determine whether the case's degree centrality was generally increasing or decreasing over time. To isolate cases that have ceased to be important, I used this linear regression data to filter the cases down to those whose downwardly sloping regression lines crossed the *x*-axis before the year 2013, indicating that the number of citations had effectively dropped to zero. This method provided a simple time-to-failure analysis that enabled me to calculate the total useful life of each failed case, defined as the year of failure minus the year the case was published. For example, if a 1993 case's regression line crossed the *x*-intercept in 1998, indicating citations to the case essentially ceased that year, it's life span would be five years. Below is an example of one such case.



This case was moderately cited after it was published, but has since failed. The trend lines are polynomial regression lines of degrees 1 through 5.

The above figure shows the yearly change in indegree centrality scores for 1978 CanLII 368, a British Columbia Court of Appeal case considering whether a municipality had a duty to repair a highway pothole that caused injury. The case was cited with moderate frequency in the five years following its publication, but hasn't been cited at all since the early nineties.

According to the regression line, this case failed (so to speak) at 2003, following about ten years of inactivity. In contrast, consider the next figure, depicting the yearly centrality scores of 1978 CanLII 11, a strongly trending Supreme Court of Canada case describing in detail the test for finding duplicity among the charges of a criminal information.



Citations to this case have increased exponentially over time

10.2. FINDINGS

To find the age at which a significant number of cases cease to be important, I calculated this life span value for all cases with negative slope that failed prior to 2013, grouped them by database, and took the arithmetic mean of the values for each group. The average life span of a case in each database is shown below.

Database ID	Average life span of cases (in years)
csc-scc	49.3
ntca	16.3
ntsc	13.1
nttc	13.0
bcca	12.0
abca	11.4
pescad	10.2
bcsc	10.0
qcca	9.9
nlca	9.4
cci-tcc	9.3
nlsctd	9.0
nsca	8.5
abqb	8.2
fct	8.2
pesctd	7.8
ykca	7.4
nbca	6.8
oncj	6.5
skca	6.4

Database ID	Average life span of cases (in years)
nssf	6.2
nbqb	6.0
skqb	5.7
mbca	5.6
bcpc	5.6
fca	5.1
nssc	5.0
nlpc	4.9
onca	4.9
yksc	4.8
qccs	4.7
mbqb	4.5
abpc	4.2
nuca	4.2
onsc	4.0
peipc	4.0
qccq	3.9
skpc	3.8
nsfc	3.5
nbpc	3.4
yktc	3.4
mbpc	3.3
nspc	3.2
nucj	3.0
onscdc	2.8

Note that in addition to the expected result of the appellate court decisions having longer average life spans, there are some curious entries toward the top of the list, including the lower courts of the Northwest Territories, Prince Edward Island, Newfoundland, and Nova Scotia, among others. This shows that the lower courts of smaller jurisdictions issue decisions that linger on for the same lengths of time as the appellate court decisions in much larger jurisdictions like Alberta and British Columbia. These numbers support the theory that decisions in less populous jurisdictions survive longer, probably because the overall volume of newly issued law is much lower in those jurisdictions, leading courts and practitioners to cite older decisions with greater frequency and duration.

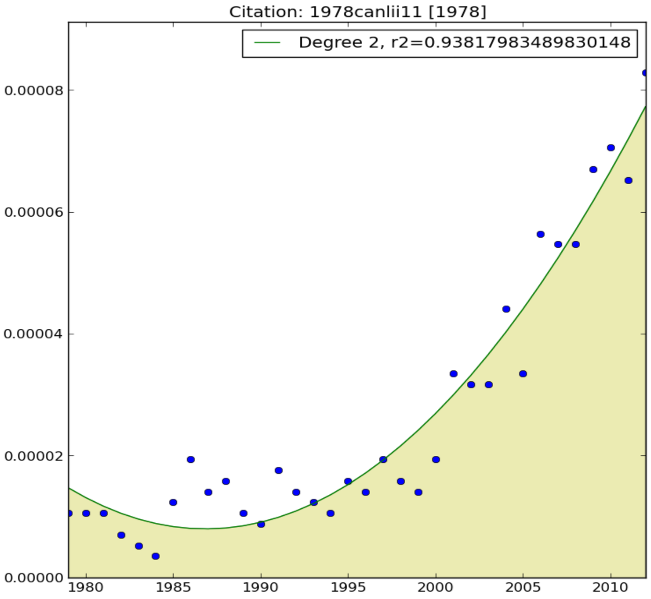
10.3. Shortcomings in Methodology

A potential shortcoming of the time-to-failure analysis outlined above is the use of a linear predictive model. The linear model is less costly to implement during exploratory analysis, but may oversimplify trends in the data and most likely has inferior predictive capabilities to other, more established models used in time-to-failure analysis, such as Weibull distributions. A similar analysis using Weibull distributions would be more complicated, but might be significantly more accurate at modeling failure and therefore presents an appealing line of future inquiry.

11. Identifying Cases that Continue to Be Important over Time

11.1. Methodology

To determine which cases continue to be important over time, I again used the year-by-year indegree centrality scores for each case computed during the network analysis phase. This time, I queried the database for cases with upwardly sloping regression lines, restricting the selected cases to those whose indegree centrality scores have generally trended upwards over time. To eliminate cases with positively trending centrality but a comparatively short lifespan, I filtered out any cases with a lifespan shorter than 15 years, the average time-to-failure across all databases. Finally, to isolate the strongest cases for the purpose of identifying any common characteristics, I sorted the cases in order of importance by computing the area underneath each regression line.



The total area under the indegree centrality curve for 1978 CanLII 11 provides a workable way to model its overall credibility versus other cases

Using this method, if two cases had an identical slope and centrality score at the time they were published, but one had remained strong for twice as long as the other, that inequality would be reflected in the absolute area beneath each case's regression line and would provide a reasonable means of distinguishing between the two.

11.2. Findings

For decisions of the Canada Supreme Court, the percentage of cases that continue to be important despite the passage of time is 18.7%, where importance is defined as a positively trending pattern of citation for a period of at least 15 years. For all other jurisdictions, the percentage of cases that remain important over time is less than 4%. The specific values for each database ID are shown in the table below.

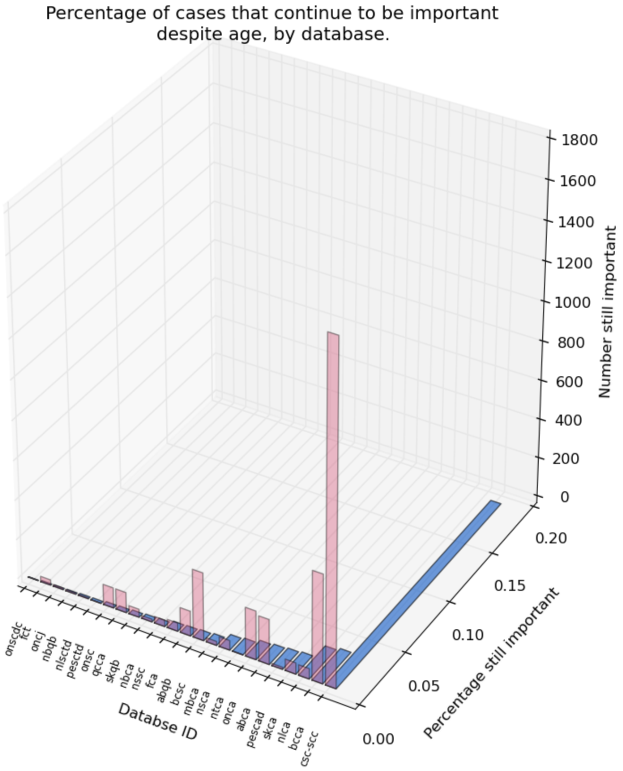
Database ID	% Important Despite Passage of Time
csc-scc	0.187
bcca	0.030
nlca	0.028
skca	0.017
pescad	0.016
abca	0.016
onca	0.015
ntca	0.012
nsca	0.011
mbca	0.007
bcsc	0.007
abqb	0.006
fca	0.006
nssc	0.004
nbca	0.004
skqb	0.003
qcca	0.003
onsc	0.002
pesctd	0.001

nlsctd	0.001
nbqb	0.001
oncj	0.001
fct	0.001
onscdc	0.001

These findings may appear to conflict with the finding that the average time-to-failure of Supreme Court cases is 50 years. The difference is attributable to the two different selections of cases used to compute the numbers: this section computes the percentage of cases that are currently positive versus all other cases, and the percentage is a relatively small percentage of the whole, whereas the previous section averaged the life spans of all previously failed cases. Though somewhat confusing, this section is still consistent with the life span observations made in the previous section.

11.3. Common Characteristics of Important Cases

Below is a plot to help visualize this information across each database. The pink vertical bars indicate the total number of cases from each database that have remained important despite the passage of time. The blue bars projected across the bottom of the chart reflect this same number, but as a percentage of the total number of cases in each database.



Number (pink) and percentage (blue) of cases per database that have remained important despite the passage of time.

The chart suggests that a higher percentage of appellate court decisions remain important over time. The higher percentage of important cases originates from the Supreme Court, unsurprisingly, followed by appeal courts in British Columbia, Newfoundland and Labrador, Saskatchewan, Prince Edward Island, Alberta, Ontario, Northwest Territories, Nova Scotia, and Manitoba. Ordering the databases by percentage of important cases allows us to obtain a sensible comparison across databases by controlling for factors like disparities in database coverage and differences in population size that affect the absolute numbers of important decisions. Looking up the individual cases with the highest areas under their indegree centrality regression lines is interesting. The top case establishes standards for determining whether an erroneous jury charge in a criminal trial is grounds for reversal (1987 CanLII 67). The second establishes the standard for obtaining a stay of execution of a judgment (1994 CanLII 117). Another is a pivotal case considering a criminal defendant's invocation of the right to counsel during interrogation (1987 CanLII 67). Many of these enduring cases appear to be cited for uncontroversial legal rules that are procedural or at least quasi-procedural in nature. For reference, the top 100 are listed in the appendix.

12. Conclusion

The indegree centrality and PageRank scores of caselaw within CanLII's database collections are effective predictors of how frequently those cases will be viewed on CanLII's website. Simple exploratory analysis of indegree citation over time versus database coverage may provide insight into the citation norms and practices unique to each jurisdiction. Furthermore, plotting the network ranking of a case over time and determining the slope and *x*-intercept of its overall trend can yield useful insights into how long cases continue to be cited before falling into relative disuse. Similar techniques can also help pinpoint the most influential cases, which are frequently cited for procedural or quasi-procedural points of law. More accurate measurements of the time-to-failure of cases might be obtained using Weibull distributions instead of a linear model, and more informative conclusions might be drawn about jurisdiction-specific citation norms if a greater breadth of historical data were available and more facets of citation practice were examined, such as the frequency with which courts in each jurisdiction cite cases decided in other jurisdictions.

13. Appendix

13.1. Top 100 Cases That Continue to be Cited Over Time

1991canlii93	1993canlii116	1985canlii46
1994canlii117	1989canlii13	1990canlii104
1987canlii84	1990canlii45	1988canlii8
1996canlii230	1978canlii11	1990canlii55
1990canlii90	1994canlii39	1985canlii23
1992canlii25	1986canlii17	1992canlii50
1996canlii191	1984canlii21	1974canlii168
1984canlii33	1990canlii70	1989canlii77
1996canlii183	1994canlii28	1985canlii47
1987canlii17	1990canlii118	1980canlii21
1991canlii45	1997canlii324	1987canlii67
1992canlii89	1990canlii32	1990canlii95
1993canlii105	1989canlii93	1996canlii229
1995canlii51	1986canlii29	1990canlii138
1997canlii384	1987canlii25	1993canlii70
1997canlii319	1987canlii74	1992canlii31
1979canlii8	1982canlii24	1993canlii3011
1993canlii34	1993canlii2939	1982canlii22
1986canlii46	1990canlii29	1993canlii286
1990canlii52	1989canlii87	1992canlii2417
1995canlii47	1989canlii2728	

1995canlii150	1982canlii20	
1976canlii2	1995canlii3498	
1974canlii14	1996canlii255	
1987canlii79	1997canlii342	
1990canlii77	1997canlii389	
1990canlii125	1979canlii23	
1980canlii22	1994canlii2570	
1994canlii127	1997canlii345	
1994canlii80	1988canlii80	
1989canlii123	1993canlii3379	
1994canlii64	1985canlii74	
1995canlii59	1994canlii65	
1992canlii56	1993canlii3375	
1993canlii126	1979canlii10	
1985canlii29	1984canlii25	
1993canlii146	1988canlii73	
1978canlii1	1995canlii72	
1995canlii108	1993canlii68	
1989canlii34	1975canlii146	

14. Glossary

adjacent

Two nodes constituting an edge are "adjacent" in the network.

context-free grammar

A recursive set of rules for tokenizing input and assembling the tokens into a parse tree.

nodes

A node is an element in a graph that can connected to other nodes.

edges

An edge is a set of two nodes and represents a connection between them.

degree

The degree of a node in a network is the number of connections it has to other nodes.

degree distribution

The degree distribution of a network is the probability distribution of the degree of all nodes in the network.

scale-free network

A network is scale-free if its degree distribution follows a power law.

random network

A network is random if its degree distribution is normal, or shaped like a bell curve.

degree centrality

The number of edges for a given node.

in-degree centrality

The number in-bound edges for a given node.

out-degree centrality

The number of out-bound edges for a given node

eigenvector centrality

A centrality measure that assigns relative centrality scores to all nodes in a network in a way to accords greater weight to connections to high scoring nodes.

betweenness centrality

A centrality measure equal to the number of shortest paths from all nodes to all others that pass through that node.

authority score

A measure of the extent to which a case is cited by other cases that tend to cite authoritative cases.

hub score

A measure of the extent to which a case tends to cite authoritative cases.

multiplicity

Occurs when one case cites another multiple times.

outliers

Cases cited many times within a corpus, potentially skewing calculations that fail to account for multiplicity.

References

Chandler, S.J. (2005), *The Network Structure of Supreme Court Jurisprudence*, Public Law and Legal Theory Series, University of Houston Law Center No. 2005-W-01. Available at: http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID742065_code254274.pdf?abstractid=742065&mirid=1. (accessed 10th December, 2013)

Clark, T.S. and Lauderdale, B.E. (2012), *The Genealogy of Law*, Political Analysis, Vol. 20 No. 3 pp. 330-331. Print. Available at http://userwww.service.emory.edu/~tclark7/genealogy.pdf. (accessed 10th December, 2013)

Cross, F.B. et al. (2010), *Citations in the U.S. Supreme Court: an Empirical Study of Their Use and Significance*, University Illinois Law Review, No. 2 pp. 489-575. Print. Available at: http://illinoislawreview.org/wp-content/ilr-content/articles/2010/2/Cross.pdf. (accessed 10th December, 2013)

Fowler, J.H. et al. (2007), *Network Analysis and the Law: Measuring the Legal Importance of Precedents at U.S. Supreme Court*, Political Analysis, No. 15 pp. 324–346. Available at http://jhfowler.ucsd.edu/network_analysis_and_the_law.pdf (accessed 10th December, 2013)

Fowler, J.H. et al. (2008), *The Authority of Supreme Court Precedent,* Social Networks, Vol. 30, No. 1, pp. 16-30. Print. Available at http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID1008032_code646904.pdf?abstractid=1008032&mirid=1 (accessed 10th December, 2013)

Geist, A. (2009), *Using Citation Analysis Techniques For Computer-Assisted Legal Research in Continental Jurisdictions*, Graduate thesis, University of Edinburgh. Available at: http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID1397674_code1087080.pdf?abstractid=1397674&mirid=1(accessed 10th December, 2013)

Gerhardt, M.J. (2008), *The Irrepressibility of Precedent*, North Carolina Law Review, Vol. 86, No. 5, pp. 1279- 1297. Available at: http://ssrn.com/abstract=2306700 (accessed 10th December, 2013)

Lupu, T. et al. (2012), *Precedent in International Courts: A Network Analysis of Case Citations by the European Court of Human Rights*, British Journal of Political Science. Print. Available at http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2015331_code1021034.pdf?abstractid=1643839&mirid=1 (accessed 10th December, 2013)

Malmgren, S. (2011), *Towards a Theory of Jurisprudential Relevance Ranking. Using Link Analysis on EU Case Law*, Graduate thesis, Stockholm University, Chapter 3.2.1.

Smith, T.A. (2005), *The Web of Law*, San Diego Legal Studies Research Paper No. 6-11. Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=642863 (accessed 10th December, 2013)