Network analysis as an aid to legal interpretation

Can counting and drawing rules help lawyers understand the context of those rules?

By John F Fitzgerald

John F Fitzgerald *is a solicitor in private practice in NSW specialising in Administrative Law. He is am updating author of the administrative law chapters of* Thomson's Laws of Australia

Abstract

This paper presents a novel method of interpreting the context of legislation based on a branch of mathematics: Graph theory.

The subject of the study is the *Health Practitioner Regulation National Law 2009* (NSW), a statute concerned with registration and discipline of health practitioners in the state of New South Wales.The study focuses on terms defined in that statute.The purpose of the study is to demonstrate network analysis can assist lawyers better understand the context of a statute.

The study shows most members of the group of defined terms have a type of relation with one another – resulting in a structure or organisation of those terms. In addition, the study shows that because of this structure some terms are most sensitive to change and most terms are insensitive to change.

No prior knowledge of Graph theory is needed to understand the paper. All mathematical concepts are explained in non-technical language. No mathematical notation is used.

Introduction

Interpretation of large collections of rules, such as legislation or contracts, is a daily task for practising lawyers. The act of interpretation requires understanding the context of the particular text in question. Yet, how do lawyers know the context? This study proposes network analysis can be a way to know the context.

Network analysis is a quantitative research technique based on a branch of mathematics: Graph theory. Despite its name Graph theory is not applying a formula to data and plotting its results in a bar graph or a line graph. Graph theory is the mathematical study of the collection of things related in some way to one another.

The focus of network analysis is the relation between things, not the things themselves. So, in the study presented in this paper the focus is on the occurrence of one defined term in the definition of a second defined term. The other aspects of the meaning of both terms is irrelevant. The whole collection of things and the relations between them is called a “graph” or “network”.

The things are called “nodes” or “vertices”. The relations between them are called “arcs” or “edges”. If the reader can imagine a set of places linked by roads then the places AND roads are a graph. The places are the nodes. If the roads are all one-way roads then the roads are called “arcs”. If the roads are all two way roads then the roads are called “edges”. If all the roads are one-way then the relations in the network are oriented from one node to the next the network is called a “directed” network. The number of arcs or edges connected to a node is its “degree”. In a directed network the number of arcs pointing away from a nod is called its “outdegree” and the number of arcs pointing inwards is called its “indegree”.

Graph theoretic analysis produces two kinds of results: quantitative measurements and diagrams. Both kinds of output are used in this study. This paper uses a software program called *Pajek*[1] to carry out all the analyses.

This paper is not novel in using Graph theory to analyse legislation, but the use of Graph theory to analyse legislation is historically new and uncommon. Chandler 2005[2] is a study of cross-references among sections of the Uniform Commercial Code of the United States. Other graph theoretic studies of cross-references within and among legislation are listed at the bibliography at the end of this article.

This study differs to all earlier works except that of Winkels et al. in 2013[3] in taking a “bottom up” view rather than a “top down” view. This study commences from the theoretical insight that all networks can be decomposed into sets of three nodes and any links between them (triads) and proceeds to examine larger structures composed of those triads. If a biological analogy is used the previous studies are at the level of anatomy whereas this study commences at the level of molecules.

To demonstrate the proposed method the paper analyses a group of defined terms in the *Health Practitioner Regulation National Law 2009* (NSW) (the *HPRNL*).

Current research methodology

Now, lawyers discover the context of statutory language by reading around the passage in question and gaining knowledge of the legislation's history and its relation to other pieces of legislation. A lawyer in Australia might use the publically accessible AUSTLII databases of legislation and search on the word or phrase in question.

Suppose a change in legislation alters the definition of a defined term. When this happens how will this change affect all other defined terms whose definition includes the now-altered defined term? There is a possibility of a ripple effect: an alteration in the meaning of one term alters the definition of a second term which in turn alters the definition of a third term and so forth.

Running an AUSTLII search on the first altered term will tell a lawyer which other defined terms have definitions containing the altered-defined term so the lawyer could then look at them. However, to track any ripple effect the lawyer would have to run further AUSTLII searches on the further altered terms. Readers will understand the point without further elaboration: individual AUSTLII searches – the conventional research tool - can only reveal direct changes occurring to one term at a time.

The subject matter of the study: the *Health Practitioner Regulation National Law 2009* (NSW)

The subject matter of the study is a collection of defined terms in the *HPRNL*, in particular, occurrences of defined terms in the definition of other defined terms. The *HPRNL* regulates the major health care professions in the state of New South Wales: doctors, physiotherapists and others. The *HPRNL* relies on statutory definitions. They are found mainly in section 5 and Division 1 of Part 8. The statutory definitions in the HPRNL number in the hundreds.

Creation of the database

The author created the database of defined terms manually. The dataset is available for non-commercial purposes on application to the author.

Research questions and findings

The study seeks to answer two questions with a common theme: is there an underlying structure in the network? Following from this question is another: do those structures cause one word in the network to be more sensitive to change in the meaning of other words in the network?

The study finds:

1. There are two structures, here called "hierarchies"; a smaller one of 17 terms and a larger one of 75 terms; and

2. The two terms most sensitive to change are: "Complaint about a registered health practitioner" s. 144, and "Unsatisfactory professional conduct" s. 138.

The network

The network of defined terms in the *HPRNL* is a set comprised of words defined elsewhere in the set of defined terms. Like earlier network analyses of non-legal dictionaries[4], the approach taken here is to regard each of the defined terms as a node. The study follows Batagelj et al.'s approach[5] by regarding the relation where their definition includes one of the other defined terms as an arc between the nodes representing the two defined terms oriented from the defined term towards the other defined term occurring in the first term's definition.

Type of graph	Directed
Number of nodes	204
Number of arcs	339
Lowest degree	0
Maximum degree	38

Table 1: Preliminary information about the network of defined terms

The network of defined terms has terms with no degree; they are called "isolates". There are 26 isolates. The term "confer" in Schedule 7 is an example.

Apart from the 26 isolates all other terms have arcs between them. A pair of nodes with arcs or edges between them is called a "dyad". Most dyads in the network have just one arc between them: they are called "assymetric dyads". However there is one pair with arcs going in both directions: a "mutual dyad" between "appellable decision" (section 175) and "appropriate responsible tribunal" (section 175(2)).

A set of three nodes and links between them is called a "triad". The triad is an important concept in network analysis[6]. No links need to exist, or more than one link might exist. All directed networks decompose into triads of one or more types. There are 16 types of triads. When different types of triads combine they produce a structure that differs to the structure produced if triads of only one type combine.

Triads vary in their complexity. In the network of defined terms the most complex is one where the first term refers to the second, that refers to a third that refers back to the first, called a "cycle"; for example: "Month" (Schedule 7 section 12), "Calendar month" ((Schedule 7 section 12) and "Named month" (Schedule 7 section 12).

Discrete parts of a network are called "components". If the reader can imagine a component of a directed network as a set of one way roads or paths, if it is not possible to travel from any given node to any other given node because of the direction of roads and it would only be possible to do so by ignoring the direction of the roads, then the component is called a "weakly connected component". On the other hand, if the direction of roads is such that it is possible to travel from any given node to any other given node, the component is called a "strongly connected" component.

In the network of defined terms there are three weakly connected components:

- One large subnetwork of 172 terms;
- A second small subnetwork of three terms: "Executive officer" (section 41A), "Council" (section 41A) and "Relevant Council" (section 41E(2)); and
- A third small subnetwork of three terms: "Council" (Schedule 5F section 1), "Return date" (Schedule 5F section 15(3) and "Return period" (Schedule 5F section 15(3)

The cycle and the mutual dyad are the only strongly connected components.

The existence of complex structures

Preliminary

The study puts to one side some parts of the 204 term database. These parts of no further relevance are: the isolates, the two small discrete weakly components of three; and the cycle. After this process there remains a subnetwork comprised of 169 defined terms called the "169 term subnetwork". There are 331 arcs.

From simple to less simple structures

The methodology of the study uses a simple, almost elementary, approach. The approach is to look for the simplest small structure that can be combined into larger structures and then look for the simplest large structure, built up from the simple small structure, which explains the whole network under investigation.

As mentioned earlier, in network analysis the simplest small structure into which a network can be decomposed is the triad. The taxonomy of 16 types of triads can be ranked according to the complexity of the combined elements: from three isolates to a cycle where there are arcs in both directions between all three node. If we start at the least complex end of the triadic spectrum it needs to be noted the 169 term subnetwork has no isolates. This property allows us to ignore the first three triad types. The next most complex triad types are 021D, 021U and 021C. They are respectively: one node refers to both of the other two nodes but those two don't refer to each other (type 021D); one node is referred to by both of the other two nodes but those two don't refer to each (type 021U) and one node refers to the second which refers to the third (type 021C). So, combinations of these three types will be the focus of further investigation.

To investigate the population of these three triad types we subtract, from the 169 subnetwork, smaller subnetworks comprised of each of the triad types. The results are:

TRIAD TYPE	NUMBER OF TERMS IN INDUCED SUBNETWORK	NUMBER OF TERMS MISSING FROM 169 TERM SUBNETWORK
021C	152	17
021U	138	31
021D	130	39

Table 2: Comparison of triads

Table 2 shows that although all three triad types account for most of the terms in the 169 term subnetwork the type 021C triad accounts for greatest number: 152 (91%).

Pendent dyads

Seventeen terms in the 169 term subnetwork do not form part of the 152 term subnetwork made up of the type 021C triad model. Why is this so? If the reader can visualize the 17 terms they would appear as pendants hanging off the "main" body of the 152 term subnetwork. If we were to look at the terms around the term to which each pendent node is linked ("the middle term") the orientation of the arc linking the pendent term to the middle term would be the same as all other arcs linked to the middle term. Either the middle term has all arcs pointing to it, or all arcs pointing away from it. This feature causes all triads formed by the pendent term, the middle term and any other term to be either a type 021U triad or a type 021D triad; but not a type 021C triad.

Once the 17 pendent terms are trimmed from the 169 term subnetwork then the remainder, called the "152 term subnetwork", is comprised of defined terms that form part of one or more 021C type triads.

Readers should note: the 17 pendent terms will not be forgotten. Their existence remains relevant to later analysis of "betweenness centrality scores" (under the heading "Sensitivity to change").

Paths

Returning to the suggested approach of looking for progressively more complex combinations of nodes, the question is: what, after the type 021C type triad, would be the next most complex possible structure? The author suggests the answer is a sequence comprised of two or more overlapping type 021C type triads.

Table 3 below presents the frequency of this suggested model structure. Some of the terminology in the table requires explanation. In Graph theory a directed network of just one type 021C triad is called a "path": you can travel from the first to the third without repeating your steps. Paths can be a combination of any number of overlapping type 021C triads. Since, strictly speaking, one type 021C triad is a path the study will from this point refer to a combination of overlapping type 021C type triads as an "explanation path" – to distinguish this model from the widely understood jargon meaning of the word “path”.

The table uses the words “explaining” and “explained”. This usage is intended to denote the orientation of arcs in the first or last terms of a triads or explanation paths. So, a triad or explanation path is thought of as commencing with a term with a zero indegree (an “initial” or “explained” term) and finishing with term with a zero outdegree (a “terminating” or “explaining” term).

The length of a path is the number of steps from beginning to end.

FREQUENCY			COMMENT
	NUMBER INITIAL WORDS EXPLAINED BY EITHER 021C TYPE TRIADS OR AN EXPLANATION PATH	48	
	NUMBER OF TERMINATING TERMS EXPLAINING EITHER 021C TYPE TRIADS OR EXPLANATION PATHS	37	
AVERAGE			
	AVERAGE LENGTH OF PATHS	3.1150	
MINIMUM			
	LENGTH	2	A 021C TYPE TRIAD
	NUMBER	13	ACCOUNTING FOR 39 TERMS
	NUMBER TERMINATING IN AN EXPLAINING TERM	4	
	NUMBER FROM AN INITIAL EXPLAINED TERM	9	
MAXIMUM			
	LENGTH OF LONGEST PATH	15 STEPS (16 TERMS)	FROM INITIAL (EXPLAINED) TERM: “confidential information notice” S 138; TO TERMINAL (EXPLAINING) TERM: “individual” Schedule 7

Table 3: Triads and paths

These statistics reveal three insights. Firstly, the high number of subnetworks consisting of explanation paths as opposed to 021C type triads: of 48 in total, only 13 are type 021C triads. This explains the average path length figure exceeding 2; the length of a type 021C triad.

Second, 48 type 021C triads or explanation paths finish in 37 terms. This fact contradicts any notion the overall network is structured as many explained terms leading up through layers of terms to just one ultimate explaining term. Rather, there are many explaining terms themselves not explained by other terms. In addition, and crucially: 11 triads and explanation paths combine.

Finally, there is one explanation path which is very lengthy: 15 steps by comparison to the average path length of 3.115 steps. Its existence with the 36 other 021C type triads and explanation paths is puzzling: one possibility is a complex underlying structure much like the spine of a human skeleton where the 36 other triads and explanation paths are other, minor, skeletal systems connected to the major skeletal system of the spine.

Hierarchy

At this point the focus of the study becomes confirmation or contradiction of the last suggested possibility: the “skeleton” model. In Graph theory the structure next most complex after a path is called a “tree”. Trees are any structure produced by overlapping or combining paths.

To simplify further investigation it is suggested further analysis be guided by the “equivalence” concept. This should occur in a similar place within the structure. The strictest form of equivalence is structural equivalence, where nodes share identical neighbours. A less strict form of equivalence is one where terms with similar, but not identical, links to each other and other terms are similarly placed in the structure.

In the following model the equivalence relation is less strict than “structural equivalence”.

The model proposes a hierarchy-like structure where:

- The model accommodates the 48 terms with a zero indegree: these are at the base of all structures (here called “layer 1 terms”);
- The model requires of terms that have both indegree and outdegree their inward oriented arcs only come from terms in the next lower layer and its outward oriented arcs only go to terms in the next higher layer (for example all terms in layer 2 with both in and out degrees only have an inward arcs from terms in layer 1 and outward arcs only to terms in layer 3); and
- The model requires every term in higher layers with no outdegree only have their inwards oriented arcs from terms in the next lower layer (for example all terms in layer 3 with no

outdegree only have an indegree from terms in layer 2)

The author concedes this is but one type of structure called a tree in Graph theory and different views can be taken in Graph theory as to the strictness of the equivalence relation. Because it is but one type of tree the term "hierarchy" will be used here to denote the chosen model.

After the 152 term subnetwork is searched for the model structure the 152 term subnetwork decomposes into:

- A 120 term subnetwork which decomposes into:
    - o A large hierarchy accounting for 75 terms with 16 classes or layers;
    - o A second hierarchy accounting for 17 terms with 5 classes;
    - o Two single-triad components; and
    - o 11 single dyad-components; and

- 32 terms are excluded: they have no indegree, so they otherwise be allocated to layer 1, but their arcs do not link to terms in layer 2.

The 120 term subnetwork is visualized as follows:

of 32 terms to the next section: "Sensitivity to change").

The larger of the two hierarchies (75 terms), explaining the term with label C24 "confidential information notice" (section 138) has a complex internal structure. The "spinal" structure, hypothesised above, appears. At different nodes going down the longest path relatively shorter paths join: for example

- at term with label P29 "protected report" (section 138);

- Terms with labels U02, U03, U04 and U05: respectively "unsatisfactory professional conduct" (section 138), unsatisfactory professional conduct of registered health practitioner (section 139B); unsatisfactory professional conduct of medical practitioners (section 139C) and unsatisfactory professional conduct of a pharmacist (section 139D).

The longest path represents the points where smaller paths join. Its significance is this union of smaller structures into a larger structure.

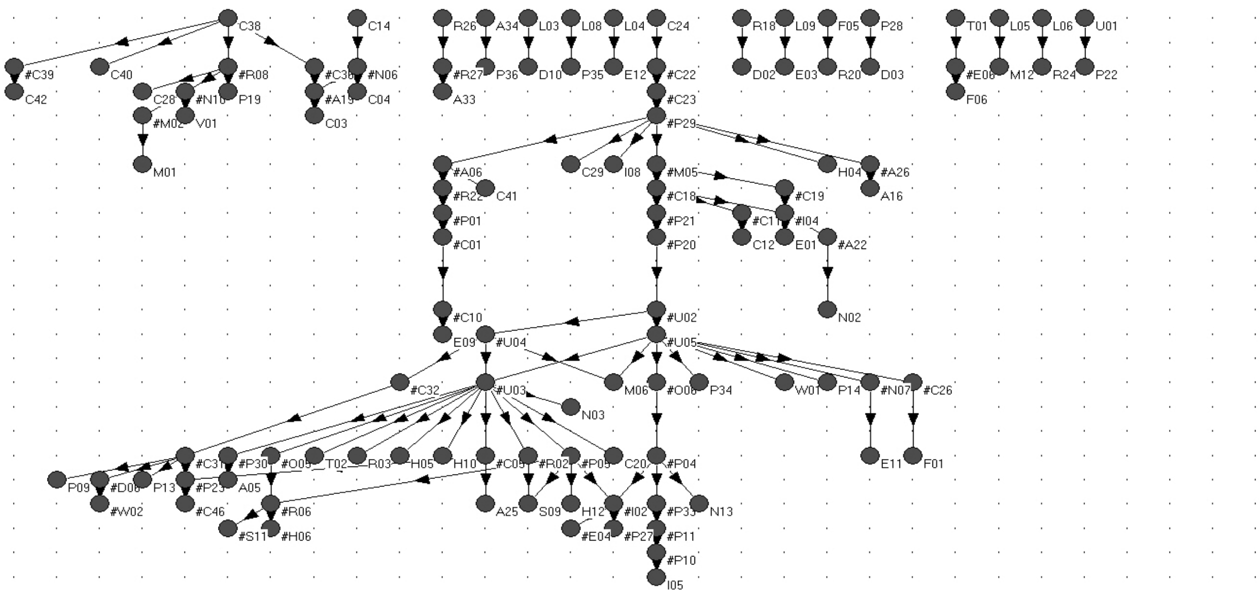

Figure 1: 120 term subnetwork hierarchies

Discussion

The principal features of this analysis are:

- the structures depicted in Figure 1: the two hierarchies and the many smaller components; and

- the group of 32 terms not depicted in the diagram (we postpone discussion of the group

The existence of the two hierarchies answers the primary research question in the affirmative: many terms (92 in number) can be organized according to a hierarchy structure based on an overall or meta-organizing principle: terms' relations with others are located by the "most explained/least explaining to least explained/most explaining" idea.

Sensitivity to change

In this section the paper discusses whether it is possible to systematically quantify sensitivity to change in the context of a statutory provision?

Network analysis suggests one way to answer the question is "betweenness centrality":

Of all the explanation paths traversing defined terms what is the defined term traversed most often as an intermediary or step along way?

Table 4 below lists the betweenness centrality scores for terms in the 172 term subnetwork:

SIZE OF NETWORK	172 TERMS	
MINIMUM SCORE	0	
MAXIMUM SCORE	.0157	
AVERAGE SCORE	.00006	
DISTRIBUTION OF SCORES		
0	103	
0 - .0053	64	
.0053 - .0105	3	
.0105 - .0157	2	
TOP 10 SCORES	0.0157	C18 Complaint about a registered health practitioner S 144
	0.0151	U02 Unsatisfactory professional conduct S 138
	0.0101	U03 unsatisfactory professional conduct of medical practitioners S 139C
	0.0070	P29 Protected report S 138
	0.0060	M05 Matter SS 139F, 145B(3)
	0.0052	U05 Unsatisfactory professional conduct of pharmacists S 139D
	0.0049	C23 Confidential information S 139A
	0.00327	C31 Court S 138
	0.00322	C19 Complaint about a student S 144A
	0.0025	R06 Registered health practitioner S 5

Table 4: Betweeness Centrality Scores 172 term network

Discussion

There is a large variation in the betweenness centrality scores. The scores divides into two groups: a smaller group of 64 with non-negligible scores and a larger number of terms of 103 with negligible scores. Further, the scores of the smaller group is unevenly distributed: for example the score of number 10 in the list of top scores above (.0025) is less than a sixth of number one (.0157). A theory explaining these facts is the existence of a structure of some terms with a “one-to-many” relation to other terms and many terms with a “one-to-one” relation to other term. In other words, there are few terms into which many explanation paths “feed” and many terms in only one path.

This theory can be visually compared to figure 2.The terms with the highest six betweenness centrality scores are located at junctions of two paths in the 75 term structure. It follows that twice as many “journeys” across that structure will cross them compared to terms around them on only one explanation path. Thus the existence of the structure and their location within the structure causes their centrality scores.

But what of terms not in the two hierarchies? Figure 2 shows 11 dyads and two triads and then there are the further 32 terms not displayed in Figure 2. The group of 32 terms would otherwise be located at the top of the diagram. They are not displayed because they are in dyads with terms not in the layer immediately below. To this should be added the 17 pendent terms in dyads excluded from the 152 term subnetwork in the first place. This is 77 terms: more than a third of the entire 172 term subnetwork. They will be called “exterior terms”.

The location of the exterior terms has the opposite effect on their centrality scores. They are all pendent on larger structures so they are not between any or many terms. Hence their betweenness centrality scores are negligible. Again, the existence of the complex structures and the location of exterior terms outside the structures causes their centrality scores.

Conclusion

This study sets out to show a graph theoretic approach to statutory interpretation can help lawyers understand the context of statutes. The study finds the most prevalent kind of structural relation among defined terms in the statute under examination, the *HPRNL* is: a type 021C triad. These combine into longer paths. These longer paths in turn combine to create two multilayered hierarchies: the smaller a hierarchy of 17 terms and 5 layers, the larger a hierarchy of 75 terms and 16 layers. In addition the study shows these structural features of the network make some terms most sensitive to change and the large number of terms insensitive to change.

[1] deNooy W, Mrvar A, Batagelj V: *Exploratory Social Network Analysis with Pajek*, Cambridge University Press (2010) (de Nooy, Mrvar and Batagelj 2010).

[2] Chandler, S. J. "*The Network Structure of the Uniform Commercial Code: It's A Small World After All*" Wolfram Technology Conference. 2005. Available at library.wolfram.com/infocenter/Conferences/5800 (Chandler 2005).

[3] Winkels, Radboud, Alexander Boer, and Ivan Plantevin. "Creating Context Networks in Dutch Legislation." *Proceedings of the 2013 Conference on Legal Knowledge and Information Systems JURIX. 2013: The Twentysixth Annual Conference*, IOS Press (2013) (Winkels and others 2013).

[4] e.g. Batagelj, V., Mrvar A., Zaveršnik M. "Network analysis of dictionaries". *Language Technologies*, T. Erjavec, J. Gros eds., Ljubljana, p. 135-142 (2002) (Batagelj and others 2002), and Blondel, Vincent D. and Senellart Pierre P. "Automatic extraction of synonyms in a dictionary.Technical Report 89", Catholic University of Louvain (2002) (Blondel and Senellart 2002).

[5] Ibid (Bantagelj et al)

[6] Ibid n. 1 (p 201)