PROXIMITY AND NEIGHBOURHOOD: USING TOPIC MODELLING TO READ THE DEVELOPMENT OF LAW IN THE HIGH COURT OF AUSTRALIA

DAVID J CARTER* AND ADEL RAHMANI**

There is significant interest in the computational analysis of legal text to enhance search capabilities, to navigate the ever-growing corpus of electronic legal texts, and to aid in the application of artificial intelligence approaches to legal work and analysis. However, the feasibility of utilising this technique to trace the emergence and evolution of particular legal doctrines or substantive areas of law over time has not been achieved so far, despite the centrality of this task to legal practice and the use of legal text. Here, we report on the application of computational methods, including topic modelling and visualisation techniques, to measure the emergence and development of particular doctrines or areas of substantive law. We do so using two case studies, firstly the Mabo *litigation regarding native title, and the rise and fall of 'proximity' as a test of foreseeability and the duty of care in tort.*

I INTRODUCTION

Legal reasoning relies on complex practices of legal reading. These legal reading practices contest and recontest what texts we read and how we read them in aid of the collective development of legal doctrine and its application. Multiple modes of reading are mobilised for these purposes. At times, legal reading requires attention to punctuation and word order, to tense or other features of language. At other times it requires attention to the selection and rejection of entire texts themselves. These and other modes of reading are part of the reading practices of the discipline of law, a discipline that demands a skilful switching between various scales at

* Dr David J Carter is a member of the Faculty of Law at the University of Technology Sydney with an appointment to the Centre for Health Economics Research and Evaluation at the UTS Business School. He is a member of the Law, Health, Justice Research Centre in the Faculty of Law and is currently a National Health and Medical Research Council ('NHMRC') Early Career Fellow (Grant ID: 1156520). The contents are solely the responsibility of the individual author and do not reflect the views of NHMRC.

** Dr Adel Rahmani is a Senior Lecturer in the School of Mathematical and Physical Sciences, Faculty of Science, University of Technology Sydney.

Acknowledgments: Part of this work was conducted on computational facilities provided by the UTS eResearch High Performance Computer Cluster. We would like to acknowledge and thank Professor James Brown our collaborator for his guidance and support of this work. We would like to particularly thank the anonymous reviewers of this article. Their careful and critical reading of our work has spurred new work to better grapple with methodological underpinnings of this form of cross-disciplinary work.

which to read a legal text; anywhere from a 'close' to 'distant' reading scale.[1]

The computational analysis of legal text provides yet another approach to legal reading; albeit one somewhat different from those practices commonly used by legal scholars and practitioners. Computational methods for engaging with and analysing text that emerge from the 'digital humanities' have advanced a nuanced and humanities-oriented engagement with these computational methods in ways that can be fruitful for the practices of legal reading. So too do they fit well within that lineage of empirical legal scholarship that proposes to engage with legal text by converting — or perhaps transforming — qualitative material into quantitative judgments,[2] particularly on a highly inclusive scale that engages a (very) large number of legal texts simultaneously. The use of computational analysis of legal text has already produced some significant and original results by analysing the entire corpus of opinions of the United States ('US') Supreme Court, or large samples of judgments handed down by other superior courts.[3]

There is still significant work to be transacted in the development of specifically *legal* approaches to the computational analysis of *legal* text. As Mireille Hildebrandt rightly locates the challenge, the stakes inherent in the reading of legal texts do not lessen when engaging text via computational means.[4] The inherent authority and the effect/consequences of a particular reading of legal text is different to literary or other texts and '[t]his raises the question of the difference between human interpretation and computational pattern-recognition and the issue of whether this matters for the meaning of law'.[5]

Not only is the challenge surrounding computational methods and their application to legal text one of the 'stakes' of legal textual interpretation. It is also, in part, a methodological one. Legal scholarship 'frequently pursues doctrinal, interpretive, and normative purposes rather than empirical ones'.[6] In so doing, the legal discipline engages in methods founded on *persuasion*, and so relying on rhetoric and attendant practices of hermeneutics and interpretation that are shared by scholars and the legal community. This includes scholarly efforts to

1 We are here referring to the concept of 'distant reading' popularised by Franco Moretti, and in circulation amongst digital humanities scholarship: see Franco Moretti, *Distant Reading* (Verso, 2013); Shawna Ross, 'In Praise of Overstating the Case: A Review of Franco Moretti, *Distant Reading* (London: Verso, 2013)' (2014) 8(1) *Digital Humanities Quarterly*.

2 Valerie P Hans, Jeffrey J Rachlinski and Emily G Owens, 'Editors' Introduction to Judgment by the Numbers: Converting Qualitative to Quantitative Judgments in Law' (2011) 8(S1) *Journal of Empirical Legal Studies* 1.

3 See below Part III for an overview of the application of topic modelling in particular. For an overview of the uses to which topic modelling has been put, see David J Carter, James Brown and Adel Rahmani, 'Reading the High Court at a Distance: Topic Modelling the Legal Subject Matter and Judicial Activity of the High Court of Australia, 1903–2015' (2016) 39(4) *University of New South Wales Law Journal* 1300, 1302–8.

4 Mireille Hildebrandt, 'The Meaning and Mining of Legal Texts' in David M Berry (ed), *Understanding Digital Humanities* (Palgrave Macmillan, 2012) 145.

5 Ibid 145.

6 Jack Goldsmith and Adrian Vermeule, 'Empirical Methodology and Legal Scholarship' (2002) 69(1) *University of Chicago Law Review* 153, 153.

> reconcile or distinguish lines of precedent displaying internal tensions; that provide conceptual analysis of the internal logic of statutes, cases, and other materials; and that provide novel readings of canonical legal sources. The best legal scholarship combines these features, fitting confused canonical materials together in a coherent way and presenting the materials in a normatively attractive light. Work in this vein contains no empirical claims in any important or contestable sense …[7]

Legal scholarship of this type relies upon interpretative and hermeneutic methodological practices based not upon its own intradisciplinary traditions regarding legal knowledge, method and their relationship to truth. Rather, 'legal scholars often write in the lawyer's style rather than in the empiricist's because they are participants in, not just students *of*, the legal system's practices'.[8] This means that legal (scholarly) methods are founded upon the elaborated rules and accepted practices of formal legal interpretation that provides normative or 'legal' reasons for adherence to them. There are, in short, legally acceptable ways of engaging and interpreting text, and legally unacceptable or at least *unpersuasive* ways of doing so. These methods are sophisticated, and are governed by the 'detailed rules of legal methodology in relation to the hierarchy of legal texts'[9] including the operation of the doctrine of precedent, 'together with the much more detailed and generally less than explicit features of the "legal art" of interpretation and argumentation within specific legal disciplines or bodies of law'.[10] These rules, taken together, aim to advance the rule of law.

This mode of legal scholarship is what we provide in this article; doctrinally-focused, attentive to the hermeneutic, interpretative and persuasive norms of legal scholarship. However, we do so by drawing upon a computational method, and so face the challenge of cross-disciplinarity and its attendant conflicts regarding, particularly, the evaluation of the suitability and success or failure of the methods we apply to legal text. Our chosen method of topic modelling is, strictly speaking, not (yet) an acceptable method for interpreting legal text according to the normatively-influenced standards and expectations of legal doctrinal scholarship. For this reason, this article is itself an attempt to work through and model a way of cross-disciplinary legal scholarship that draws upon both computational and legal

7 Ibid 155.

8 Ibid (emphasis added).

9 Peter Goodrich, *Legal Discourse: Studies in Linguistics, Rhetoric and Legal Analysis* (Macmillan, 1987) 173.

10 Ibid.

disciplines, whilst informed by the demands of legal theory and practice.[11] Given this, our aim here is to utilise computational methods to engage with legal text, in a manner that is primarily oriented towards its 'usefulness'[12] for advancing legal doctrinal scholarship; measured primarily — but not solely — by whether or not our efforts are able to advance an argument that *persuades* in its attempt to fit 'canonical materials together in a coherent way and [present] the materials in a normatively attractive light'.[13] This is what we mean when we describe specifically '*legal* approaches to the computational analysis of *legal* text' above.

In this article we pursue an examination of what computational approaches to legal text can facilitate. To do so we demonstrate the use of topic modelling and visualisation to measure the emergence and development of particular legal doctrines in the recent history of the High Court of Australia. In previous work,[14] we utilised the topic modelling technique named latent Dirichlet allocation ('LDA'),[15] a probabilistic method, which uses Bayesian inference to identify latent features from the text, and which is commonly used for these purposes. In this article we present the results of using an alternate method, which uses matrix factorisation to identify latent features from the text — non-negative matrix factorisation ('NMF')[16] — for tracing the emergence and evolution of legal doctrines or substantive areas of law. We do so using two case studies, firstly the *Mabo* litigation regarding the recognition of native title, and the rise and fall of 'proximity' as a test of foreseeability and the duty of care in tort. To do so, we first provide an overview of the High Court of Australia and its engagement with both the *Mabo* litigation and the test of proximity. Following this, we provide a brief engagement with the potential and existing uses of topic modelling and other computational methods to analyse legal text.

11 In this, we are influenced by the work of Hildebrandt, and her work at the interface between information and communication infrastructures (like written text and its associated practices) and its co-constitution of law itself: see especially Mireille Hildebrandt, *Smart Technologies and the End(s) of Law: Novel Entanglements of Law and Technology* (Edward Elgar Publishing, 2015) chs 7–8. See also Hildebrandt's discussion of this 'core thesis' in her work: Mireille Hildebrandt, 'Law *as* an Affordance: The Devil Is in the Vanishing Point(s)' (2017) 4(1) *Critical Analysis of Law* 116; Mireille Hildebrandt, 'Legal and Technological Normativity: More (and Less) than Twin Sisters' (2008) 12(3) *Techné* 169.

12 We trust that our use of the word 'usefulness' is read not as a statement of legal disciplinary hegemony over that of another discipline and its methods. Rather, we are here 'caught', as it were, in a meeting of 'neighbouring foreign sovereigns'; two disciplines, with their own practices of jurisdiction, meeting to perform legal work. Although outside of the scope of this work, the work of Dorsett and McVeigh on jurisdiction — and especially its performance and 'meeting of laws' — might be a way to think through cross-disciplinarily. See Shaunnagh Dorsett and Shaun McVeigh, *Jurisdiction* (Routledge, 2012); Shaunnagh Dorsett, *Juridical Encounters: Māori and the Colonial Courts, 1840–1852* (Auckland University Press, 2017).

13 Goldsmith and Vermeule (n 6) 155.

14 Carter, Brown and Rahmani (n 3).

15 See David M Blei, Andrew Y Ng and Michael I Jordan, 'Latent Dirichlet Allocation' (2003) 3 *Journal of Machine Learning Research* 993.

16 See Andrzej Cichocki and Anh-Huy Phan, 'Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations' (2009) 92(3) *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 708; Cédric Févotte and Jérôme Idier, 'Algorithms for Nonnegative Matrix Factorization with the β-Divergence' (2011) 23(9) *Neural Computation* 2421.

II BACKGROUND

The High Court of Australia is the most senior court of the Australian judicial system. Established by the *Australian Constitution*, the Court exercises an original jurisdiction over constitutional and other matters, whilst it holds the discretion to hear appeals as the ultimate court of appeal from other federal, state and territory courts.

The Court has operated since 1903, and has been a perennial topic for study and analysis amongst Australian legal scholars. A small but important body of scholarship has approached the work of the Court by way of quantitative or empirical means. The work of Groves and Smyth, for example, reveals the patterns and form of judicial writing during the 20th century.[17] The ongoing work of Lynch and Williams provides statistical analysis of the Court's workload and other features of its judicial decision-making in an annual series.[18] Our own work also stands in this stream, having earlier applied the method of quantitative textual analysis known as 'topic modelling' to the collective judgments of the High Court of Australia from 1903 to 2015 in order to examine the work of the Court.[19]

The High Court of Australia has been the site of some of the most high profile and significant contests of legal doctrine and practice in Australian history. Two such important contests of doctrine are the focus of our work here: the *Mabo* litigation and the rise and fall of the test of 'proximity'; two areas of law that relate, in very different ways, to the question of neighbourhood and neighbour.

These two case studies represent areas of law and developments within the Court that are not directly related in any substantive way. What they do share, however, is a history that occurs at a similar time, a shared status as important and at times controversial innovations in the recent history of the Court, and a significant body of scholarly and professional commentary and literature. We do not profess deep expertise in the substantive law, history or reception of the two case studies. Instead, we aim to utilise the case studies to develop and demonstrate the topic modelling and visualisation techniques, offering the fruits of that analysis as a prompt to further scholarship by others. Given this, the case studies were selected for their status as important areas of the Court's work, for the ability for the results of our work here to speak to and be engaged by an ongoing scholarship regarding

17 Matthew Groves and Russell Smyth, 'A Century of Judicial Style: Changing Patterns in Judgment Writing on the High Court 1903–2001' (2004) 32(2) *Federal Law Review* 255. See also Yanir Seroussi, Russell Smyth and Ingrid Zukerman, 'Ghosts from the High Court's Past: Evidence from Computational Linguistics for Dixon Ghosting for McTiernan and Rich' (2011) 34(3) *University of New South Wales Law Journal* 984. In relation to Australian supreme courts, see Russell Smyth, 'The Business of the Australian State Supreme Courts over the Course of the 20th Century' (2010) 7(1) *Journal of Empirical Legal Studies* 141.

18 Andrew Lynch and George Williams, 'The High Court on Constitutional Law: The 2015 Statistics' (2016) 39(3) *University of New South Wales Law Journal* 1161.

19 See Carter, Brown and Rahmani (n 3). This excludes special leave dispositions, transcripts and bulletins.

the two topics and for the different challenges they provide to the technique of topic modelling, namely that one is a well-defined sequence of cases related to the same facts (the *Mabo* litigation), whereas the other is a group of factually unrelated cases that are important for their development of a specific doctrinal feature (proximity). We introduce each of these case studies in the following Part.

A *The* Mabo *Litigation*

The *Mabo* litigation is a collection of decisions made by the High Court during the 1980s and 1990s that remains one of the most commented upon series of decisions in its history. The name 'Mabo' refers to Eddie Koiki Mabo who, with other Meriam men Reverend David Passi, Celuia Mapoo Salee, Sam Passi and James Rice, brought an action against the State of Queensland claiming title to the Murray Islands on behalf of the Murray Islanders. The high point of the *Mabo* litigation was the decision of *Mabo v Queensland [No 2]* ('*Mabo [No 2]*'), where the High Court held that, at common law and under certain circumstances, a form of 'native title' survived English colonisation of Australia in 1788.[20] The case held that native title may exist where a community, group or clan has had a continuing connection with land in accordance with traditional laws, and the title has not been extinguished by a valid inconsistent Crown act. In *Mabo [No 2]* the Court found that the Murray Islanders had the right as against the whole world to possession, occupation, use and enjoyment of the relevant parts of the Murray Islands.[21]

There is a vast literature on the *Mabo* litigation, its history and development.[22] The place of the doctrine of terra nullius is particularly important in the reception of the case, with a dominant interpretation being that the case overturned the doctrine. This line of reading the case holds that Australian real property/land law had claimed that Australia was terra nullius or 'land belonging to no one' and that the High Court had rejected or reversed this doctrine.[23] However, the doctrine has been found to be absent from the historical period in which it was said to have been used to structure colonial dispossession:

20 (1992) 175 CLR 1 ('*Mabo [No 2]*').

21 Ibid 76 (Brennan J, Mason CJ and McHugh J agreeing).

22 See, eg, Keon-Cohen's writing that places the decision(s) in their legal context: Bryan Keon-Cohen, 'From "Land-Related Agreements" to "Comprehensive Settlements" to "Domestic Treaties": An Inevitable Progression?' (2018) 25 *Pandora's Box* 43. See also Irene Watson's work on the subjection of Indigenous people by the Australian colonial state, including by way of terra nullius and its 'capacity to bury [Indigenous people] and its own capacity to survive and go on burying [Indigenous people]': Irene Watson, 'Buried Alive' (2002) 13(3) *Law and Critique* 253, 253. See also Tony Birch, 'The Last Refuge of the "Un-Australian"' in Timothy Neale, Crystal McKinnon and Eve Vincent (eds), *History, Power, Text: Cultural Studies and Indigenous Studies* (UTS ePress, 2014) 198; Peter H Russell, *Recognizing Aboriginal Title: The* Mabo *Case and Indigenous Resistance to English-Settler Colonialism* (University of Toronto Press, 2005).

23 See especially David Ritter, 'The "Rejection of Terra Nullius" in *Mabo*: A Critical Analysis' (1996) 18(1) *Sydney Law Review* 5; Andrew Fitzmaurice, 'The Genealogy of Terra Nullius' (2007) 38(129) *Australian Historical Studies* 1.

> When Australia was originally colonised by the Crown, neither *terra nullius* or any other legal doctrine was used to deny the recognition of traditional Aboriginal rights to land under the common law. Such a doctrinal denial would not have appeared necessary to the colonists, because the indigenous inhabitants of the colony were seen and defined by the colonists as intrinsically barbarous and without any interest in land. Thus the colonists required no legal doctrine to explain why Aboriginal people's land rights were not to be recognised under law because no doctrine was required for what was axiomatic.[24]

As Ritter puts it, the question of whether Aboriginal people held some legal right to their land was instead litigated first in 1971, and until that point, no doctrinal reason for their dispossession had been developed.[25] This is to say, despite the long and ongoing reality of dispossession, the practice of such dispossession had not arisen as a formal question of land law until late in the 20th century.

For our purposes, what is most important in this debate is to understand the role of land law/real property law as central to the 20th century engagement with indigenous rights and sovereignty — despite the relatively recent pedigree of the characterisation of that question of dispossession as one relating to land law. How this might be reflected in the development of the *Mabo* litigation across its various incarnations before the High Court, and how our method can illustrate or illuminate these changes over time is then central to our task. As we show below, the cases that make up the *Mabo* litigation in fact did transition from those centred on constitutional matters, ending with *Mabo [No 2]* itself which established native title as firmly a matter of land law. That is to say, our methods here demonstrate an accurate interpretation of the *Mabo* litigation — in so far as it accords with the broader view of their development and character.

B *Proximity*

The second focus of our work here is the principle or test of 'proximity' as the basis of the duty of care in tort. The law of negligence had long struggled to erect firm boundaries around the question of the duty of care, and for a period of time, 'proximity' became the approach by which the High Court sought to answer 'the duty question'.[26] The duty question traces its lineage to Lord Atkin's use of the 'neighbour test' to define where and when a duty of care is owed to another in the landmark case of *Donoghue v Stevenson*.[27] His Lordship wrote there:

> The rule that you are to love your neighbour becomes in law, you must not injure

24 Ritter (n 23) 6 (emphasis in original).

25 Ibid.

26 Desmond Manderson, 'The Ethics of Proximity: An Essay for William Deane' (2005) 14(2) *Griffith Law Review* 295 ('The Ethics of Proximity').

27 [1932] AC 562 ('*Donoghue v Stevenson*').

> your neighbour; and the lawyer's question, Who is my neighbour? receives a restricted reply. You must take reasonable care to avoid acts or omissions which you can reasonably foresee would be likely to injure your neighbour. Who, then, in law is my neighbour? The answer seems to be — persons who are so closely and directly affected by my act that I ought reasonably to have them in contemplation as being so affected when I am directing my mind to the acts or omissions which are called in question.[28]

With this, Lord Atkin is said to have laid the foundation of contemporary negligence law. The significance of his Lordship's approach is located in its reliance not upon separate and individual categories where a duty arises, but by way of a broad principle: whether or not it was reasonably foreseeable that the plaintiff would be injured by my act or omission.

However, the flexibility and breadth of this now principle-based approach risked drawing into the frame of 'neighbour' too many persons, particularly evident in the Australian context by way of causes of action for negligent misstatements causing pure financial loss as established in *Caltex Oil (Australia) Pty Ltd v The Dredge Willemstad*.[29] In short, the concern was that the principle may open the 'floodgates', leading in the well-known words of US Chief Justice Cardozo, to 'liability in an indeterminate amount for an indeterminate time to an indeterminate class'.[30]

It was in that context that 'proximity' arose as a limit on responsibility. In *Jaensch v Coffey*,[31] Deane J — a High Court Justice who was to become associated particularly with proximity — interpreted the 'neighbour principle' of Lord Atkin in *Donoghue v Stevenson* by emphasising the notion of proximity as an existing, but underutilised limb of the test for a duty of care. Deane J argued that this approach was to focus on the question posed by Lord Atkin of '[w]ho, then, in law is my neighbour?'[32] Lord Atkin's response was that 'I ought reasonably to have them in contemplation as being so affected when I am directing my mind to the acts or omissions which are called in question'.[33] Deane J formulated proximity in *Jaensch v Coffey* as engaging in an evaluation of both the closeness of the relationship and of the legal consequences of judging that evaluation as sufficiently or insufficiently proximate,[34] emphasising that 'proximity could not be confined to physical proximity, but could include "circumstantial" and

28 Ibid 580.

29 (1976) 136 CLR 529.

30 *Ultramares Corporation v Touche*, 174 NE 441, 444 (NY Ct App, 1931).

31 (1984) 155 CLR 549 ('*Jaensch v Coffey*').

32 Ibid 580.

33 *Donoghue v Stevenson* (n 27) 580.

34 Prue Vines, 'The Needle in the Haystack: Principle in the Duty of Care in Negligence' (2000) 23(2) *University of New South Wales Law Journal* 35, 43.

"causal" proximity'.[35]

For his Honour, and eventually the High Court,[36] 'proximity' was to function as an overriding control on the test of reasonable foresight as the determinant for a duty of care in negligence. Within that framing, the proximity of relationship between the parties is an overriding control on the test of reasonable foreseeability, and the relationship of proximity will determine the categories of case in which the common law will recognise a duty to take reasonable care to avoid foreseeable injury to another.[37]

Proximity was received by some as a 'radical and controversial jurisprudential development that led to innovation after innovation in the Court's judgments'.[38] So too was it one which commentators have criticised on many occasions for its alleged vagueness. An exchange between Hayne and Bell JJ with Senior Counsel for the respondent, transcribed from proceedings in *Brookfield Multiplex Ltd v Owners Corporation Strata Plan 61288*,[39] provides perhaps some sense of the High Court's own more recent view of the doctrine's clarity and some of the tradition of critical reception of it:

> **BELL J:** If you take vulnerability out of the equation, in relation to a tortious duty respecting pure economic loss, what feature do you rely on, apart from foreseeability to establish a duty?
>
> **MR CORSARO:** Proximity, reliance.
>
> **BELL J:** What does proximity mean?
>
> **MR CORSARO:** It means the fact that
>
> **HAYNE J:** Answer in not more than three hours, write in your number at the head of each sheet, Mr Corsaro. Commence writing.[40]

Despite — or perhaps due to — the sense of proximity's vagueness, its rise has

35 Ibid.

36 Des Butler, 'Managing Liability for Bystander Psychiatric Injury in a Post-*Hill v Van Erp* Environment' (1997) 13 *Queensland University of Technology Law Journal* 152, 153, citing *Cook v Cook* (1986) 162 CLR 376, *Australian Safeway Stores Pty Ltd v Zaluzna* (1987) 162 CLR 479, *Hawkins v Clayton* (1988) 164 CLR 539, *Gala v Preston* (1991) 172 CLR 243 ('*Gala*'), *Burnie Port Authority v General Jones Pty Ltd* (1994) 179 CLR 520 ('*Burnie Port*') and *Bryan v Maloney* (1995) 182 CLR 609.

37 Prue Vines, 'Proximity as Principle or Category: Nervous Shock in Australia and England' (1993) 16(2) *University of New South Wales Law Journal* 458.

38 Desmond Manderson, *Proximity, Levinas, and the Soul of Law* (McGill-Queen's University Press, 2006) 14. See also Desmond Manderson (ed), *Essays on Levinas and Law: A Mosaic* (Palgrave Macmillan, 2009) ('*Essays on Levinas and Law*'); Desmond Manderson, '"Current Legal Maxims in Which the Word Neighbour Occurs": Levinas and the Law of Torts' in Desmond Manderson (ed), *Essays on Levinas and Law: A Mosaic* (Palgrave Macmillan, 2009) 111 ('Current Legal Maxims').

39 (2014) 254 CLR 185.

40 Transcript of Proceedings, *Brookfield Multiplex Ltd v Owners Corporation Strata Plan 61288* [2014] HCATrans 126, 56.

also been cause for a rich legal scholarship in private law. Desmond Manderson's sustained reflections on proximity argue that the Court, through proximity, was 'groping towards a new idea of the nature of and justification for our ideas of responsibility'.[41] In making this claim, Manderson's work draws together the High Court's own reasoning with the philosophy of Emmanuel Levinas, arguing that the Court's own interpretation came very close to mobilising the principle of proximity in a manner akin to Levinas:

> As opposed to some of the Court's own language, however, in which proximity is understood as a limit on responsibility or as just another (and frustratingly vague) rule … proximity explains and justifies responsibility. It also indicates precisely the direction in which the court ought to go in determining its nature and boundaries.[42]

Manderson's engagement with proximity elevates the Court's own oftentimes halting and apologetic use of proximity, demonstrating instead its ethical and doctrinal coherence alongside its power, a set of features so often lacking in alternate solutions to the duty question:

> It is, rather, a question of proximity. What is lacking [in alternate constructions of the duty question], as Levinas shows us, is the intimacy of the relationship between plaintiff and defendant. This is not just a question of knowledge or names or numbers but of a relationship that specifically connects the vulnerability of the one to the capacity of the other to effect the situation.[43]

Despite these advantages, the Court alongside many other commentators has written of the 'fall' of proximity. Kirby J concluded in 1998 in *Pyrenees Shire Council v Day* that 'it is tolerably clear that proximity's reign in this Court, at least as a universal identifier of the existence of a duty of care at common law, has come to an end'.[44] Moreover, in a 2011 joint judgment, six members of the High Court concluded that '[t]he demise of proximity as a useful informing principle … is now complete'.[45] There is, however, a dissenting voice on this question in the writing of Andrew Robertson, who claims in recent work that the 'concept of proximity', just not this particular *term*, has been utilised by the Court right up to the present day.[46] According to Robertson's analysis, proximity's disavowal may well be an incomplete repudiation, leaving the status of the proximity principle misunderstood by many commentators and perhaps by the Court itself; leaving

41 Manderson, 'The Ethics of Proximity' (n 26) 299.

42 Ibid 298 (emphasis omitted).

43 Manderson, *Proximity, Levinas, and the Soul of Law* (n 38) 158.

44 (1998) 192 CLR 330, 414 ('*Pyrenees*').

45 *Miller v Miller* (2011) 242 CLR 446, 468 [59] (French CJ, Gummow, Hayne, Crennan, Kiefel and Bell JJ) ('*Miller*').

46 Andrew Robertson, 'Proximity: Divergence and Unity' in Andrew Robertson and Michael Tilbury (eds), *Divergences in Private Law* (Hart Publishing, 2016) 9.

open the possibility that the proximity principle may well be alive and thus providing significant justification for analysis and reassessment of proximity in the work of the High Court. Given the social and doctrinal import of both the *Mabo* litigation and of proximity in the recent life of the High Court, our role here is to report one way by which this kind of legal scholarship and doctrinal analysis might be supported.[47]

III TOPIC MODELLING

Given the overview of the substantive legal terrain upon which we wish to apply topic modelling provided immediately above, in this Part we provide a brief engagement with the potential and existing uses of topic modelling as a way of introducing and contextualising our own legal methodological experimentation below.

Topic modelling is a family of machine learning and statistical approaches that can extract *latent* (ie abstract) topics from a corpus of text. A topic model is produced by application of quantitative textual analysis, a process used to identify repeated occurrences of collections of words. In perhaps the clearest description of the technique for lawyers and legal scholars, Megan Brett describes the process of topic modelling by analogy with manual text analysis by use of multiple 'highlighters' where differently coloured highlighters are used to mark different themes or topics on a page of printed text.[48] How this plays out in a topic model is that each topic within a topic model is represented as a series of words (or more properly 'tokens') ordered by the likelihood of their appearing in documents belonging — in part — to that topic. In our earlier work topic modelling the High Court of Australia, for example, one topic was characterised by the tokens 'constitution', 'parliament', 'government', 'regulation', 'territory', 'legislative', 'federal', 'constitutional' whilst another 'trial', 'offence', 'criminal', 'jury', 'accused', 'crime', 'police', 'prosecution'.[49] One topic here describing matters of public law and particularly that of the Commonwealth or federal powers whilst another clearly relating to the criminal law.

Given this set of tokens, a challenge of 'naming' the topic remains. Lisa M Rhody notes that the result of the topic modelling process is that topics are produced, but that this topic modelling process 'does not produce names for the topics it discovers or sort words with an understanding of what words *mean*'.[50] Rather, topic modelling is 'an *ex post facto* method for challenging our assumptions about

47 Consistent with our earlier work on the topic, we see the form of analysis afforded by techniques like topic modelling as a method for the study of legal texts and practices that might fruitfully complement other methods of legal scholarship and practice: see Carter, Brown and Rahmani (n 3).

48 Megan R Brett, 'Topic Modeling: A Basic Introduction' (2012) 2(1) *Journal of Digital Humanities*.

49 Carter, Brown and Rahmani (n 3) 1316.

50 Lisa M Rhody, 'Topic Modeling and Figurative Language' (2012) 2(1) *Journal of Digital Humanities* (emphasis in original).

natural language data'.[51]

Topic modelling presents a relatively new technique for law and legal textual analysis. It is, at present, used infrequently in legal scholarship. However, there is a growing body of work that utilises topic modelling both in an applied mode and more experimentally, where colleagues focus on the development of the underlying technique and approach to modelling. Scholars have recently utilised topic modelling to study the Austrian Supreme Administrative Court decisions[52] and the *combined* case corpora of the US Supreme Court and appellate courts.[53] Livermore, Riddell and Rockmore have developed an approach to studying the 'agenda formation' of the US Supreme Court, utilising topic modelling and other approaches,[54] and more recently have presented a study of the increasing distinctiveness of the US Supreme Court's opinions from the 'judicial genre'.[55] In more doctrinally-focused work, Macey and Mitts have constructed and utilised a dataset of 9,380 US cases that reference either corporate veil-piercing or other forms of disregarding the corporate form,[56] against which they fit a model to study corporate veil-piercing in that jurisdiction.[57] Joshua Mitts has developed a model to track the rise of new regulatory issues to support 'predictive regulation' efforts,[58] whilst the work of Daniel Taylor Young uses topic modelling to test 'empirically' a theory of US constitutional change developed by Bruce Ackerman.[59]

In our previous article, we used topic modelling to 'read' the decisions of the High Court of Australia at different 'distances';[60] an approach engaged with the textual practices developed by those working in the digital humanities. This was not the first time topic modelling had been applied to this legal textual corpus.[61] However,

51 Ibid (emphasis in original).

52 Ingo Feinerer and Kurt Hornik, 'Text Mining of Supreme Administrative Court Jurisdictions' in Christine Preisach et al (eds), *Data Analysis, Machine Learning and Applications* (Springer, 2008) 569.

53 Michael A Livermore, Allen Riddell and Daniel Rockmore, 'A Topic Model Approach to Studying Agenda Formation for the US Supreme Court' (Virginia Law and Economics Research Paper No 2, University of Virginia School of Law, 10 July 2015).

54 See their draft paper, Michael A Livermore, Allen B Riddell and Daniel Rockmore, 'Agenda Formation and the US Supreme Court: A Topic Model Approach' (29 February 2016).

55 Michael A Livermore, Allen B Riddell and Daniel N Rockmore, 'The Supreme Court and the Judicial Genre' (2017) 59(4) *Arizona Law Review* 837.

56 Jonathan Macey and Joshua Mitts, 'Finding Order in the Morass: The Three Real Justifications for Piercing the Corporate Veil' (2014) 100(1) *Cornell Law Review* 99, 141.

57 Ibid 149.

58 Joshua Mitts, 'Predictive Regulation' (SSRN Scholarly Paper, 27 June 2014) <ssrn.com/abstract=2411816>.

59 Daniel Taylor Young, 'How Do You Measure a Constitutional Moment? Using Algorithmic Topic Modelling to Evaluate Bruce Ackerman's Theory of Constitutional Change' (2013) 122(7) *Yale Law Journal* 1990.

60 Carter, Brown and Rahmani (n 3). For a recent discussion of these concepts and their expression in various techniques: see S Jänicke et al, 'On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges' (Conference Paper, Eurographics Conference on Visualisation, 2015).

61 Seroussi, Smyth and Zukerman (n 17); Yanir Seroussi, Ingrid Zukerman and Fabian Bohnert, 'Authorship Attribution with Topic Models' (2014) 40(2) *Computational Linguistics* 269; Yanir Seroussi, Ingrid Zukerman and Fabian Bohnert, 'Authorship Attribution with Latent Dirichlet Allocation' (Conference Paper, Conference on Computational Natural Language Learning, 23–4 June 2011) 181.

it was the first application that attempted a cross-disciplinary application of the method aiming to deliver relatively traditional legal doctrinal, interpretive, and normative outcomes of its application. By applying topic modelling within the frame of 'distant reading', our work provided an invitation for analysis of language use by uncovering latent structures in the Court's own repertoire of legal concepts. In the 100-topic model fit to the Court's output, for example, the five leading topics with which the model classified *Mabo v Queensland [No 1]* ('*Mabo [No 1]*')[62] were:

Table 1: Leading topics for *Mabo [No 1]*: 100-topic model

Topic	Words/Tokens	Coverage
96	Political | native | government | freedom | people	25%
41	Land | title | crown | mining | grant	23%
29	Territory | water | northern | area | aboriginal	8%
58	Council | queensland | committee | privy | privy_council	6%
89	Paragraph | clause | specified | requirement | description	5%

Intriguingly, two topics — topic 41 and 29 — were found to speak to 'land' and 'territory' respectively. In a topic model with fewer topics, both may have been expressed by a single topic speaking to a unitary concept of 'land' and the legal practices associated with land. However, in this 100-topic model, the collection of tokens/words demonstrate how, for the Court, topics 41 and 29 represent quite distinct topics. The topic model was able to discern, describe, and differentiate between 'land' (topic 41), as related to words (ie concepts and practices) such as 'title', 'crown', 'mining' and 'grant', and 'territory', which it understands in relation to 'water', 'northern', 'area' and 'aboriginal'.

The implication we drew there was that 'the Court itself understands and utilises two distinct topics in the text of their judgments, marking out two distinct practices'.[63] At the time, it seemed to us that one of these topics described the Court's engagement with the practices of classifying, controlling and using 'land' by way of the colonial legal technologies of ownership ('title', 'crown'), that necessitated/facilitated the control/exclusion of others ('title', 'grant') and generated specific purposes such as 'mining' or 'grant[ing]'.[64] The second topic seemed to us to speak to the Court's engagement with 'territory' — a largely spatialised concept accompanied by terms that, by contrast to the first topic, described parts of the earth's surface as inert and not characterised by 'use or control'.[65] There were, for example, no verbs in the topic list. Of most relevance to

62 (1988) 166 CLR 186 ('*Mabo [No 1]*').

63 Carter, Brown and Rahmani (n 3) 1334.

64 Ibid.

65 Ibid.

our present analysis, the single legal or natural subject associated with this second topic was 'Aboriginal'. By contrast, the first topic associated 'land' and all of the verbs and legal technologies with the legal and natural subjects of the 'Crown' and those engaged in 'grant[ing]' or 'mining' and who possess or claim 'title'.

Rendering visible these topics within the High Court's judgments was an invitation to engage in further analysis. For one, the contrast — in this 100-topic model at least — between legal technologies and verbs being associated with only the 'land' topic and then spatialised terms and no verbs with the 'territory' topic raised for us the question of how and why these characteristics came to be so divided. As part of the history and ongoing practices of Aboriginal dispossession practiced in and through law in Australia, the association of 'Aboriginal' as the single natural and legal subject within the key terms for both topics was particularly challenging.

To further this line of analysis, we engaged in improvement of our method of topic modelling and as reported here, began to overlay both time and part of the Court's own citational network. The following Part describes this method and is application, followed by its results.

IV METHODS: NON-NEGATIVE MATRIX FACTORISATION

Whilst most uses of topic modelling legal text have utilised LDA,[66] in this article we use an alternate, matrix factorisation approach — NMF[67] — of which we illustrate the use in tracing the emergence and evolution of legal doctrines and substantive areas of law within the collected reasons for judgment handed down by the High Court of Australia ('judgments').

In our earlier application of topic modelling, we also engaged with the entire corpus of judgments handed down by the High Court of Australia. Our focus in that context was, however, different to that which we pursue here. In that earlier work, we utilised topic modelling in a more descriptive mode, first providing a new and unique view of the judicial activity and legal subject matter before the High Court throughout its history (the 'work' of the Court). In this work we were successfully able to develop and apply topic modelling as a 'taxonomy of

66 Including our own work.

67 For an introduction to NMF, see Daniel Jurafsky and James H Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Pearson, 2nd ed, 2009); Nicolas Gillis, 'The Why and How of Nonnegative Matrix Factorization' in Johan AK Suykens, Marco Signoretto and Andreas Argyriou (eds), *Regularization, Optimization, Kernels, and Support Vector Machines* (CRC Press, 2015) 257; Daniel D Lee and H Sebastian Seung, 'Learning the Parts of Objects by Non-Negative Matrix Factorization' (1999) 401(6755) *Nature* 788; Wei Xu, Xin Liu and Yihong Gong, 'Document Clustering Based on Non-Negative Matrix Factorization' (Conference Paper, International ACM SIGIR Conference on Research and Development in Information Retrieval, 28 July – 1 August 2003) 267. See also the work of Pandey and Mohler, who have applied NMF to criminological data: Ritika Pandey and George O Mohler, 'Evaluation of Crime Topic Models: Topic Coherence vs Spatial Crime Concentration' (Conference Paper, IEEE International Conference on Intelligence and Security Informatics, 9–11 November 2018) 76.

practice', testing the models and then showing how the work of the Court had changed over time, showing how what we termed 'trial process for criminal matters' and 'government action in relation to immigration' have experienced a marked increase in their dominance of the Court's overall legal subject matter, while 'trade licensing, regulation and IP', 'company financial flows' and 'land contracts and agreements' have seen a downward trend.[68] Our earlier work was focused on the 'work' of the Court and its change over time. The work presented here, by contrast, focuses upon a finer-grained question; that of the development and change of doctrine and legal subject matters within the Court's corpus.

At this point we believe it is helpful to provide an illustration of the fundamental nature of the NMF method we apply here by way of a 'toy model': the opening paragraph of Charles Dickens' *A Tale of Two Cities*. While further technical details regarding our application of our computational methods are contained in an electronic online appendix/repository to this article,[69] we recommend that the interested reader engage with the abundant literature on these techniques and their application alongside our explanation here.[70]

Consider a corpus of four 'documents' from that opening paragraph written by Dickens:[71]

> Document 0: it was the best of times it was the worst of times it was the age of wisdom it was the age of foolishness
>
> Document 1: it was the epoch of belief it was the epoch of incredulity
>
> Document 2: it was the season of light it was the season of darkness
>
> Document 3: it was the spring of hope it was the winter of despair

For analysis, we need to convert the documents into a numerical form. A common way to do that is to transform each document into a 'bag-of-words', by counting how many times each unique word appears in the entire vocabulary of all documents, while ignoring word ordering. There are 20 unique words in this collection of documents. Represented as a table, columns represent the (alphabetically ordered) words in our vocabulary, with a row for each document (Table 2):

68 Carter, Brown and Rahmani (n 3) 1321.

69 Adel Rahmani, 'Repository for the "Proximity and Neighbourhood" Article by David Carter & Adel Rahmani', *GitHub* (Web Page) <github.com/adelr/hca-proximity>.

70 For a selection of that literature, see above n 67.

71 For simplicity we have already transformed the text to lower case and stripped punctuation symbols.

Table 2: Bag-of-words representation of our toy corpus

	age	belief	best	darkness	despair	epoch	foolishness	hope	incredulity	it	light	of	season	spring	the	times	was	winter	wisdom	worst
doc0	2	0	1	0	0	0	1	0	0	4	0	4	0	0	4	2	4	0	1	1
doc1	0	1	0	0	0	2	0	0	1	2	0	2	0	0	2	0	2	0	0	0
doc2	0	0	0	1	0	0	0	0	0	2	1	2	2	0	2	0	2	0	0	0
doc3	0	0	0	0	1	0	0	1	0	2	0	2	0	1	2	0	2	1	0	0

Table 3: TF-IDF representation of our corpus

	age	belief	best	darkness	despair	epoch	foolishness	hope	incredulity	it	light	of	season	spring	the	times	was	winter	wisdom	worst
doc0	3.83	0.00	1.92	0.00	0.00	0.00	1.92	0.00	0.00	4.0	0.00	4.0	0.00	0.00	4.0	3.83	4.0	0.00	1.92	1.92
doc1	0.00	1.92	0.00	0.00	0.00	3.83	0.00	0.00	1.92	2.0	0.00	2.0	0.00	0.00	2.0	0.00	2.0	0.00	0.00	0.00
doc2	0.00	0.00	0.00	1.92	0.00	0.00	0.00	0.00	0.00	2.0	1.92	2.0	3.83	0.00	2.0	0.00	2.0	0.00	0.00	0.00
doc3	0.00	0.00	0.00	0.00	1.92	0.00	0.00	1.92	0.00	2.0	0.00	2.0	0.00	1.92	2.0	0.00	2.0	1.92	0.00	0.00

Table 4: Normalised TF-IDF representation of the corpus

	age	belief	best	darkness	despair	epoch	foolishness	hope	incredulity	it	light	of	season	spring	the	times	was	winter	wisdom	worst
doc0	0.37	0.00	0.18	0.00	0.00	0.00	0.18	0.00	0.00	0.38	0.00	0.38	0.00	0.00	0.38	0.37	0.38	0.00	0.18	0.18
doc1	0.00	0.31	0.00	0.00	0.00	0.62	0.00	0.00	0.31	0.32	0.00	0.32	0.00	0.00	0.32	0.00	0.32	0.00	0.00	0.00
doc2	0.00	0.00	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.32	0.31	0.32	0.62	0.00	0.32	0.00	0.32	0.00	0.00	0.00
doc3	0.00	0.00	0.00	0.00	0.35	0.00	0.00	0.35	0.00	0.36	0.00	0.36	0.00	0.35	0.36	0.00	0.36	0.35	0.00	0.00

Each cell in Table 2 reports the number of times a unique word appears in a particular document. This is the result of replacing the underlying documents with a numerical representation (a row of numbers, or a vector). Importantly, the vector (or row of numbers) associated with each document does not depend on whether it shares a word with other documents in the corpus.

With this processing complete, a range of numerical analyses can be completed. In our earlier work, we utilised this bag of words representation of the High Court's collected judgments by applying LDA; that is, we trained a statistical model to extract some latent features (the topics) from the 'text' as represented by this numerical summary.

However, we can take this initial representation of the corpus further by taking into account how common words are across the entire corpus. Ideally, we'd like to find topics that will allow us to discriminate between the documents. Therefore, words that are so common that they occur in most, or all of the documents in our corpus are of limited usefulness, and their influence should be downplayed.

The previous table tells us about the *term frequency* of each word in each document. By contrast, the *document frequency* of a word tells us how many documents in the corpus contain that word. Multiplying the *term frequency* with a function of the inverse of the *document frequency* (a logarithmic transformation is typically applied to soften the rescaling effect on the most frequent words), results in the term-frequency-inverse-document-frequency measure ('TF-IDF'), which captures not only how often the word appears in each document but also reduces the weight of words that are frequent across the corpus.

Applied to our 'toy corpus' of text taken from Dickens' *A Tale of Two Cities*, something like the results shown in Table 3 emerge. Notice how words like 'age' or 'belief' which appear in only one document see their weights increase, whereas the weights of words like 'it' or 'of' which are found in all the documents remain the same. In other words, those words like 'it' that occur frequently across the corpus of documents have their weights reduced in comparison to those other words like 'belief', whose appearance in few documents causes their weight to increase.

Given this result, one final refinement is made in order to account for the reality that longer documents are more likely to contain reoccurrences of words due simply to their word count being greater than other documents. A process of normalisation is undertaken to account for this, a process that yields the results of Table 4. This normalisation process causes the weight for a word like 'age' — which occurs twice in our longest document (document 0) — to be assigned a weight of comparable magnitude to 'belief' which occurs only once in the shorter document 1. In other words, a relatively unique word, occurring in a few documents is weighted comparatively greater than a word that might occur many times across many documents.

Given these processes, Table 4 represents the final numerical representation of our corpus. This may be fruitfully compared with the earlier results in Table 2 which simply represented count data; namely, the number of times particular words were contained in each document that makes up the toy corpus.

A more detailed discussion of the TF-IDF process can be found on the online electronic appendix/repository associated with this work.[72]

A *Pre-Processing of the High Court of Australia Corpus*

Our High Court corpus consists of 7,585 decisions published by the High Court of Australia between 1903 and 2017. The data was sourced from a combination of the High Court of Australia's own repository, Australasian Legal Information Institute and Jade. Each document in our corpus is one of the collected reasons for judgment handed down by the High Court of Australia (which we have referred to for ease of expression as a 'judgment').[73]

While a more detailed discussion of the technical aspect relating to the natural language processing used to extract topics from a collection of documents can be found on the online repository associated with this work,[74] we provide some basic information below.

Pre-processing of the documents in the corpus includes the following steps:

1. We filter out commonly occurring words in the English language (stop words), as well as words common to our corpus (such as the Justices' names);
2. We restrict our vocabulary to words of at least three characters;
3. We restrict our vocabulary to words which appear in at least five judgments;
4. We restrict our vocabulary to words which appear in no more than half of the judgments (that is, very 'common' words are excluded);
5. We lemmatise the words to reduce them to their base form.

Our computations were performed using the Python programming language, using the spaCy natural language processing library,[75] and the scikit-learn

72 Rahmani (n 69).

73 The nomenclature of judgments, decisions and other terms can be quite complex — and suffers from a looseness in usage in legal scholarship and practice. Technically speaking, what is often referred to as a 'judgment' of the High Court is in fact the collected reasons for judgment handed down by particular judges of the High Court. Very often these reasons for judgment are written by individuals, however, they may also be jointly authored or an individual judge may simply agree with the reasons provided by another, adding some commentary of their own.

74 Rahmani (n 69).

75 'Industrial-Strength Natural Language Processing', *SpaCy* (Web Page) <spacy.io>.

machine learning library.[76]

B *Evaluating the Topic Models*

There are a variety of methods by which a topic model may be evaluated. These include both quantitative forms of evaluation, and qualitative forms of evaluation.

In traditional unsupervised machine clustering methods (of which topic modelling in an instance), the quality of the result is often evaluated by use of a quantitative, computed metric (eg coherence scores).[77] This is because in the absence of a ground truth, one must find a way to determine which values of the method's parameters produce the optimal final result. This principle does not translate easily into the context of law because the output of the machine learning method is not the 'final result'. In fact it is the starting point of an analysis made by a human expert, which must always be conducted and interpreted in light of accepted legal methodology. As described above in the introduction to this article, this is partially a pragmatic move on our part — prioritising accepted legal methods for the evaluation of textual interpretation is a form of deference to the truly novel nature of computational methods for performing or supporting legal analysis, especially in the Australian jurisdiction. However, beyond pragmatics alone, our choice is founded upon the normative implications of differing methods of legal textual interpretation. In short, some forms of evaluation of legal textual interpretation bring with them authorisation by law and the legal community, others not. For this reason, in this application of topic models, our use of them and our evaluation of their affordances are integrated with (currently) accepted modes of legal analysis and evaluation of such analysis. What this looks like in practice is that the 'quality' of the result is ultimately linked not to the topics themselves or a measure immanent to them, but rather to the ability of these methods to support legal analysis.[78]

For these reasons, in this study we favoured the application of evaluative methods that rely on human judgment and, in particular, those that rely on a trained human interpreter. We did this in a variety of ways. Whilst we did calculate a measure of

76 Fabian Pedregosa et al, 'Scikit-Learn: Machine Learning in Python' (2011) 12 *Journal of Machine Learning Research* 2825.

77 Keith Stevens et al, 'Exploring Topic Coherence over Many Models and Many Topics' (Conference Paper, Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 12–14 July 2012) 952. See the description of coherence scores relating to the topic model presented here in 'Appendix: A Note on the Number of Topics': Rahmani (n 69). See in application: Pandey and Mohler (n 67). We thank one of our anonymous reviewers for their insightful commentary regarding coherence scores.

78 An assessment of the quality of legal analysis is, naturally, a nuanced and communal practice that takes place over a lengthy period of time and cannot be made here.

topic coherence,[79] we applied 'eye-balling models' including the 'Top 10 Words' methods and systematic, sampled reading of cases associated with a topic.[80] This was followed by a form of extrinsic evaluation, that is, evaluating a topic model by way of its ability to perform a task. The extrinsic evaluation is the most significant and important evaluation we undertook. Our aim, as described in the introduction to this article, was to seek to contribute to the development of doctrinal scholarship regarding these two landmark areas of doctrine developed by the High Court of Australia by undertaking, specifically, the task of tracing the development of the test of proximity in tort and the *Mabo* litigation over time. Using a computational method to assist in this task is highly experimental and novel for legal methods, and so, our evaluation of success or failure is tied to the ability of these methods to provide something 'useful' for advancing more legal doctrinal scholarship.

V RESULTS

We fit two topic models to the corpus of judgments of the High Court. The first, a 10-topic model and the second a 20-topic model. This is a common approach to topic modelling of text, where controlling the 'scale' at which we 'read' a particular corpus allows for an exploration of the 'usefulness' of a particular number of topics in a model. The number of topics within a particular topic model is dictated solely the user. Any number of topics may be selected, and for a purpose of human interpretability like ours, there is no strict rule nor reason outside of 'usefulness' that may be applied to guide the choice between say a topic model that consists of 10 versus one that consists of 20 topics. That being said, the number of topics generated is a critical issue; as the number of topics rises, the risk of generating a model that is not useful for human interpretation rises, while at the same time so too does the potential for a more nuanced and fine-grained topic model emerge. For this reason, there is a balancing exercise involved. A low number of topics will provide a view of the underlying material at a higher level of generality than a larger

79 See Rahmani (n 69). We thank one of our reviewers for their pressing of this question. Whilst our views did not fully cohere, the discussion has highlighted the fundamental work that is required in this, a cross-disciplinary application of computational methods for law — given law's own methods, the normative nature of its work, the question of 'authorised' and 'unauthorised' modes of analysis of law and, finally, the discomfort law itself has with an extra-disciplinary determination of 'quality' (ie the presentation, interpretation and certainly determination of legal textual material by computational disciplinary methods).

80 As to the top 10 words most associated with each topic, it is a challenge to find an appropriate number of words for a human to judge and to formulate a 'name' for a topic. This can be influenced by topic cardinality for example: Jey Han Lau and Timothy Baldwin, 'The Sensitivity of Topic Coherence Evaluation to Topic Cardinality' (Conference Paper, Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 12–17 June 2016) 483. Each judgment identified as part of either the *Mabo* litigation or those identified as part of the canon of proximity cases was read in order to (re)acquaint the reader with the nature of those judgments and the nature of the small corpus that those judgments now formed. For assessment of the 10- and 20-topic models, and for the work (below) regarding topic hierarchies, headnotes and brief scanning/reading of the full text of judgments was undertaken starting with the judgment that was most identified with a topic, until 'saturation' was reached. As described above, this approach was operationalised in a manner that was subordinated to the disciplinary norms of legal doctrinal research, and, primarily to the use of these topic models, namely, to advance a persuasive form of legal doctrinal analysis.

number of topics. This is similar to viewing Australian legal materials through the lens of the Priestley 11 taxonomy of Australian law (a higher level of generality, because of a lower number of topics) versus a more detailed taxonomy that might be made up of the subsidiary areas of law. However, unlike the taxonomical structure of the Priestley 11, a particular topic presented by one topic model, does not 'split' into subsidiary topics when a topic model containing a larger number of topics is computed (ie a topic we might label as 'criminal law' in a 10-topic model might not split into 'offences against the person' and 'offences against property' in a topic model with a larger number of topics). Rather, each topic model, despite its differing level of granularity, should be considered sui generis.

Testing various topic models for their usefulness and accuracy in relation to the underlying text and the intended use of the topic model is the preferred approach to resolving this challenge. Within this context, our goal was to fit a set of topics to the collected judgments of the High Court of Australia in a manner that provided sufficient granularity to facilitate our objective of tracing the emergence and change of particular doctrines, to tell us something of their character, and to guide further analysis by other means. As with most topic modelling, the number of topics needed to be large enough to capture some of the more important nuances of the work of the Court, but not too large for a reader to grasp at a glance.[81] Topic modelling is performed in other contexts with far larger numbers of topics, especially where the topics themselves will not be engaged primarily by way of human interpretation as we do here. The advantage of a larger number of topics in a particular model is that there is potential to gain a *more* nuanced view than with a lower number of topics, for example in producing a form of document summary as we and others have done in earlier work.[82] In this article, we explore the results of a 10- and 20-topic model, noting that greater nuance — but more difficulty for human interpretation — might be achieved by utilising a larger number of topics.[83]

A *Ten-Topic Model*

We start with a 10-topic model (labelled from 0 to 9) trained using a vocabulary of 29,117 tokens (unique words) for the judgments of the High Court of Australia included in our corpus.

These 10 topics (0–9) are best described by reference to the weights associated with each token (ie word) that are associated with the topic. Given this, the

81 For a deeper consideration of this issue, see Eleanor Rosch, 'Principles of Categorization' in Allan Collins and Edward E Smith (eds), *Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence* (Morgan Kaufmann, 1998) 311.

82 Carter, Brown and Rahmani (n 3).

83 As we have done in earlier work using topic models with 100 topics: Carter, Brown and Rahmani (n 3). See also other efforts using a similar number of topics, for example Greg Leibon et al, 'Bending the Law: Geometric Tools for Quantifying Influence in the Multinetwork of Legal Opinions' (2018) 26(2) *Artificial Intelligence and Law* 145.

clearest representation of a topic is that created by selecting the top few (say 10) tokens (ie words) with the largest weight for each topic and then ordering them by decreasing weight.[84] We show the results of this process in Table 5.

A comparison between this topic model shown in Table 5 and the 10-topic model from our previous analysis (Table 6) demonstrates strong similarities. This similarity exists despite the fact that the pre-processing steps, the machine learning approach (NMF versus LDA), and the corpus (decisions from 2016–17 are included in this NMF analysis but were not in our previous work) are different as between the two topic modelling efforts.

Having established topics utilising NMF, we proceeded to classify each document (judgment) using these 10 topics. To do so, we normalise the weights for each topic so that they together add to one for each judgment. This produces a way of easily reviewing the relative importance or weight of each of the 10 topics for a particular judgment — expressed as a percentage. Having done so, we extracted a 'ranking' of topics for each judgment. This process provides a way of reviewing classifications of each judgment by review of the most dominant topics associated with the document. Given this data for each individual document/judgment, we can show the relative importance of each of the 10 topics for the entire corpus of High Court judgments included in our analysis (Figure 1).[85]

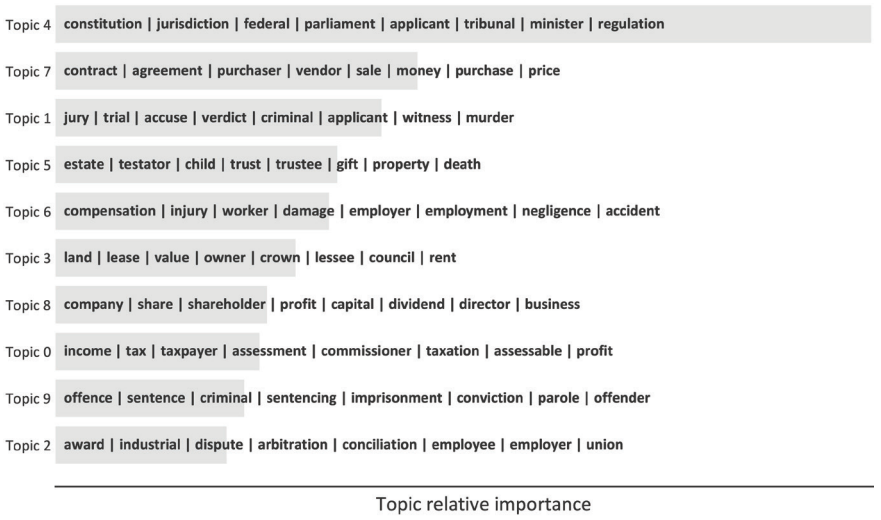

Figure 1: Topic relative importance for the 10-topic model. The longer the bar, the larger the proportion of decisions in the corpus that are associated with the topic. For each topic we also show the top eight tokens, ranked by decreasing importance from left to right.

84 Note that the order of the topics themselves is arbitrary. That is, there is no meaning that can be attributed to whether a topic is labelled as Topic 2 or Topic 8.

85 Rahmani (n 69).

Table 5: Ten-topic model

	token0	token1	token2	token3	token4	token5	token6	token7	token8	token9
Topic0	income	tax	taxpayer	assessment	commissioner	taxation	assessable	profit	deduction	taxable
Topic1	jury	trial	accuse	verdict	criminal	applicant	witness	murder	complainant	police
Topic2	award	industrial	dispute	arbitration	conciliation	employee	employer	union	organization	industry
Topic3	land	lease	value	owner	crown	lessee	council	rent	title	property
Topic4	constitution	jurisdiction	federal	parliament	applicant	tribunal	minister	regulation	government	judicial
Topic5	estate	testator	child	trust	trustee	gift	property	death	deceased	wife
Topic6	compensation	injury	worker	damage	employer	employment	negligence	accident	loss	incapacity
Topic7	contract	agreement	purchaser	vendor	sale	money	purchase	price	payment	clause
Topic8	company	share	shareholder	profit	capital	dividend	director	business	asset	value
Topic9	offence	sentence	criminal	sentencing	imprisonment	conviction	parole	offender	charge	convict

Table 6: Earlier 10-topic model utilising LDA[86]

Topic	Words	Topic Label
0	'minister', 'tribunal', 'australian', 'review', 'protection', 'conduct', 'submission', 'convention	Government action in relation to immigration
1	'child', 'trust', 'property', 'trustee', 'estate', 'wife', 'death', 'husband'	Beneficiaries' rights to property in an estate
2	'trial', 'offence', 'criminal', 'jury', 'accused', 'crime', 'police', 'prosecution'	Trial process for criminal matters
3	'trade', 'licence', 'board', 'mark', 'sale', 'patent', 'price', 'business'	Trade licensing, regulation and IP
4	'damage', 'negligence', 'injury', 'loss', 'liability', 'case', 'reasonable', 'trial'	Damage to persons through injury
5	'company', 'income', 'share', 'commissioner', 'money', 'assessment', 'business', 'payment'	Company financial flows
6	'jurisdiction', 'federal', 'judicial', 'applicant', 'hearing', 'officer', 'federal_court', 'document'	Jurisdictional divisions and actions
7	'award', 'employee', 'dispute', 'employer', 'industrial', 'employment', 'compensation', 'work'	Employment entitlements and disputes
8	'land', 'contract', 'agreement', 'property', 'lease', 'title', 'sale', 'owner'	Land contracts and agreements
9	'constitution', 'parliament', 'government', 'regulation', 'territory', 'legislative', 'federal', 'constitutional'	Constitutional actors and relationships

86 Carter, Brown and Rahmani (n 3) 1316.

Because the topics are represented as vectors of weights over the vocabulary, we are able to use a distance measure to compute how similar any two topics are. Figure 2 shows a heatmap of the cosine distance between topics as between zero and one. For simplicity the topics are labelled by their leading word. Darker colours in the heatmap indicate more similar topics (smaller distance between the topics), whilst the dendrogram placed above the heatmap shows the hierarchical clustering of topics based on this distance (using complete linkage). Note that the distances are all quite large due to the dimensionality of our problem (size of our vocabulary).

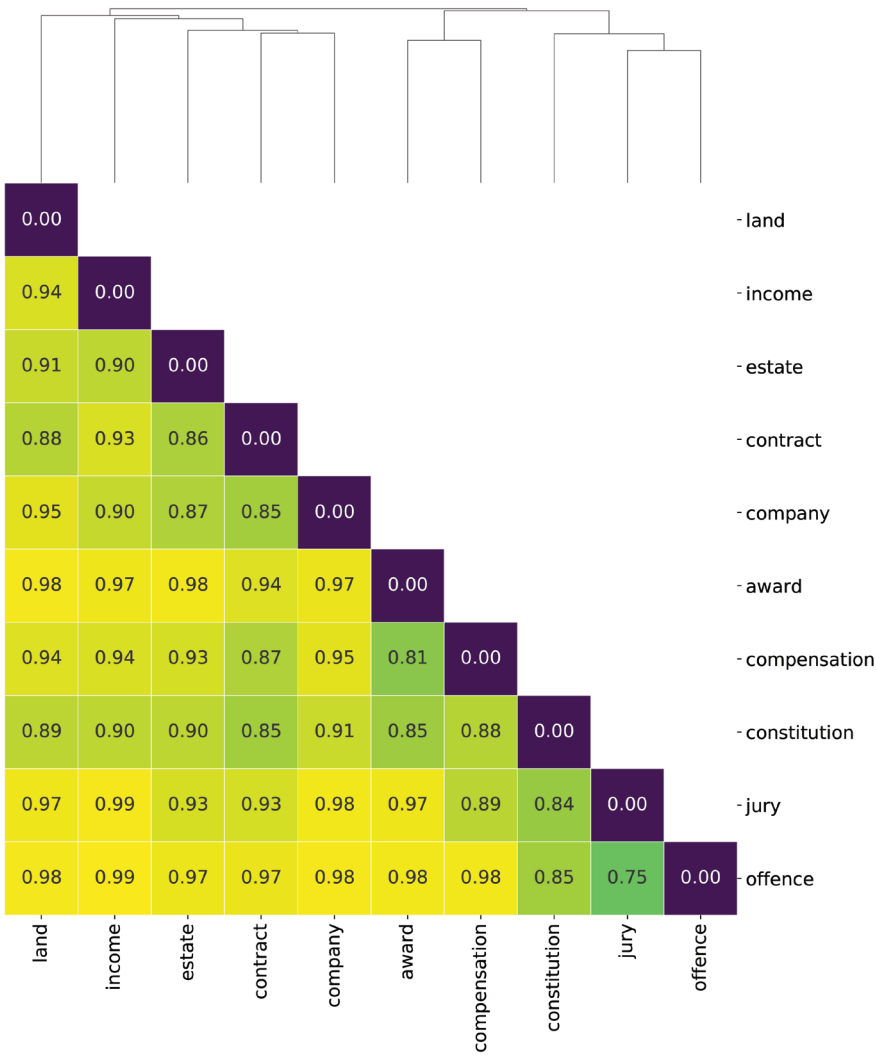

Figure 2: Heatmap of the pairwise cosine distance between the topics

There is a significant amount of information displayed in Figure 2. To illustrate its use by way of an example, note how the dendrogram suggests a splitting of the topics into three main groups. The group to the left of the diagram includes 'company', 'contract', 'estate', 'income', and 'land'. All matters of what are described in Australian legal tradition as 'private law'. The middle group clusters 'award' and 'compensation', both topics that deal with compensation, either by way of an award of damages or an industrial award (that sets wages and other conditions within the Australian industrial relations system) and compensation which relates in this case to the compensation for workplace injuries through the Australian workers' compensation system. Finally, the last group in the dendrogram is comprised of 'constitution', 'jury', and 'offence'. These final topics are all matters of 'public law'. Given this division, the dendrogram in Figure 2 highlights the underlying division of public and private law — a fundamental or foundational division used within the Australian legal system.[87] Interestingly, it separates from this primary division of public and private law the law relating to the governance of employment and workplace accident compensation. Why this is so is ultimately related to the linguistic differences (word uses) found by the model between decisions relating to these areas of law. In this three-cluster model, is it that the law of industrial relations including workers compensation is in fact a hybrid of both public and private law? Given Australia's particular tradition of industrial relations regulation nestles the concept of the (private law) contract for employment within a much larger system of government (ie public) arbitration, industrial awards, tribunals and similar, this seems a reasonable hypothesis to explore further at another time.

Figure 3 shows a clearer, rotated version of the dendrogram plotted above the heatmap. The threshold parameter (represented by the vertical dashed line in the Figure) denotes the distance below which topics are gathered into a cluster. For the threshold represented on the Figure, the topics are clustered into three groups (the three clusters mentioned above). Starting from the leftmost side of Figure 3 (distance of 0), each topic (represented by a horizontal line) is its own cluster. As we increase the distance threshold (that is, as we move towards the rightmost side of Figure 3), topics progressively cluster together. For instance, the first two topics to merge are 'offence' and 'jury'. This means that of the 10 topics these two topics are the most similar in terms of their vocabulary word weights in the NMF decomposition. This clustering makes sense, as they both relate to the broader field of criminal law and procedure. As we move to the right on the Figure, the next two topics to form a cluster are 'compensation' and 'award'. Then 'constitution' is clustered into the first 'criminal law' cluster to join together what might be understood as the 'public law' topics within the 10-topic model.

87 As it is also a feature of other common law jurisdictions worldwide.

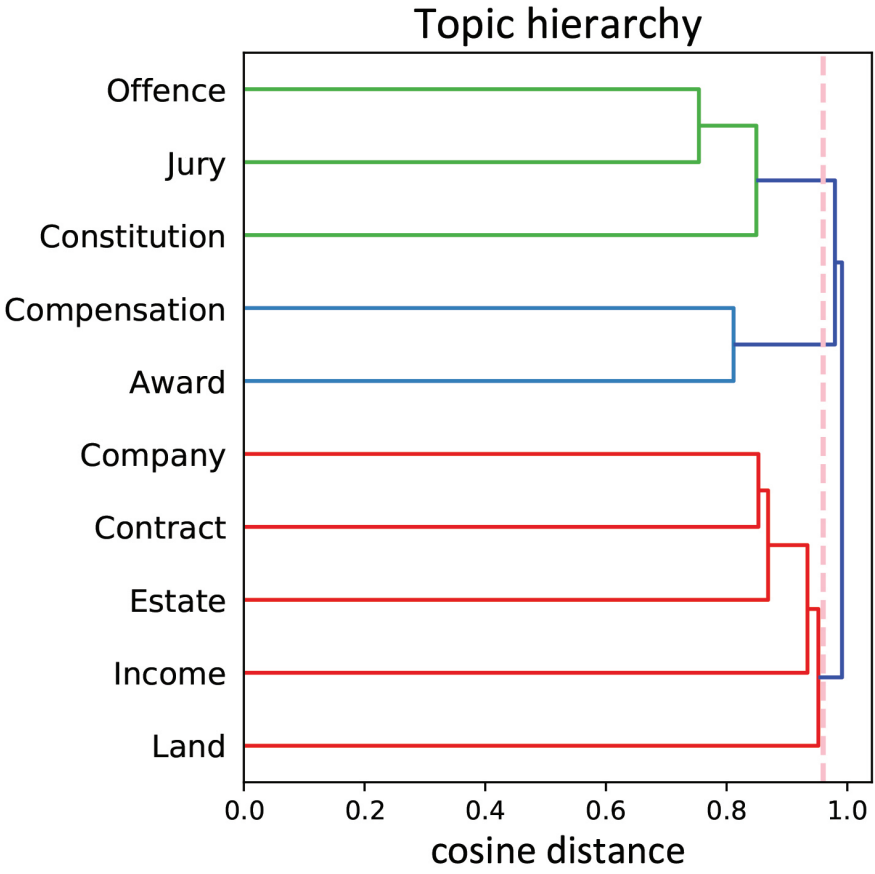

Figure 3: Hierarchical clustering of the topics (cosine distance)

If we cluster the topics into simply two clusters, then the result mirrors more completely the traditional public–private law divide firstly of the topics associated with public law, 'jury', 'award', 'constitution', 'compensation', 'offence', and then into a second cluster consisting of topics associated with private law, 'income', 'land', 'estate', 'contract', 'company'. A four-cluster model, on the other hand, divides topics by their relationship to each other into a first cluster of 'income', 'estate', 'contract', 'company', a second cluster of 'jury', 'constitution', 'offence', a third cluster of 'award', 'compensation' and finally, a cluster that consists simply of 'land'.[88] The clustering of topics makes some intuitive sense based on the subject matter that underlies the various topics. However, as is often the case with these techniques, what this might mean for our current taxonomy of the 'Priestley 11' areas of Australian law or for the accuracy and dominance of conceptions of private versus public law as dichotomous is a prompt for further analysis.

88 For further details see Rahmani (n 69).

While the topics by themselves cast an interesting light onto the work of the Court, and potentially the Australian legal system itself, we use them here primarily to advance analysis of specific legal questions, concepts and doctrines. As a first exercise we consider three cases at the heart of the *Mabo* litigation:

- *Mabo v Queensland*;[89]
- *Mabo [No 1]*; and
- *Mabo [No 2]*.

In short, the *Mabo* litigation provided an engagement with questions of the ongoing colonial and post-colonial dispossession of Aboriginal and Torres Strait Islander peoples. It was — as a whole — framed as a matter of (colonial) land law. Despite this question of land rights being the foundation of the litigation, the cases themselves cover a relatively diverse range of legal ground. For example, *Mabo v Queensland* is a short judgment made by a single judge of the High Court and relates to a decision regarding which court a hearing should be remitted to. *Mabo [No 1]* dealt with the question of the attempted extinguishment of native title rights (at that time not confirmed to have existed or in what form) by the Queensland Government by way of the enactment of the *Queensland Coast Islands Declaratory Act 1985* (Qld) which aimed to retrospectively extinguish any such rights. The case resolved a question of the validity of the Act, argued on a range of grounds. It thus did not deal with the underlying question of land and native title, its focus instead being that of inconsistency as between the Queensland (state) law and Commonwealth provisions in the *Racial Discrimination Act 1975* (Cth).[90] Finally, *Mabo [No 2]* produces a lengthy document made up of multiple judgments from multiple judges while dealing with the substantive matter that the *Mabo* litigation is known to have engaged regarding the survival of native title.

For each case we can compute a topic allocation, normalise the topic weights with respect to the dominant one, and visualise the result in a series of bar charts where, for simplicity, each topic is referred to by its dominant word (ie the word with the highest weight in the NMF representation of that topic) (Figure 4).

89 (1986) 64 ALR 1.

90 Where there is an inconsistency between Commonwealth and state law, the Commonwealth law prevails under provisions of the *Australian Constitution*.

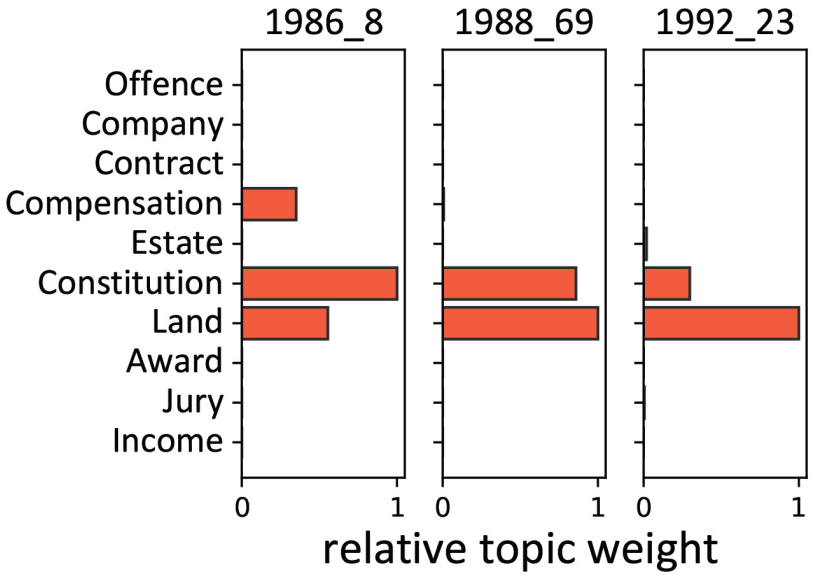

Figure 4: Topics for the *Mabo* cases. The weights are normalised so that the leading topic has a weight of 1.

A visualisation of our topic modelling of these three cases can interact with existing legal knowledge about the progress of the *Mabo* litigation in a variety of ways. Reading the three charts from left to right is to read the chronological development of the *Mabo* litigation through these three key cases. Read in that way, the topical evolution of the cases is clear. The topic model identifies that the first *Mabo* case in 1986 had mainly a constitutional focus, with some elements of land and compensation. However, between 1986 and 1988, and even more so between 1988 and 1992, the focus of these key cases shifted away from constitutional and compensation matters, towards one becoming predominantly conceptualised as an issue of 'land'. This reading accords with the meaning and focus of these cases when read by a trained interpreter. In this way, this visualisation technique is able to operate as a form of confirmation of existing legal knowledge, as a form of 'triangulating' evidence for that reading of the case law, or alternatively, as a prompt to a reading of the cases to test and examine how and why the topic model might appear in this way.

All three approaches outlined here are an *integration* of the output of the topic model into more traditional legal analysis methods. All three advance the ends of legal analysis and scholarship in some way. Reading the *Mabo* litigation through these types of analyses and visualisation allow us to quantify the broad legal subject matter of each case, the character of a case and of the variance of these features over time that might otherwise be quite difficult to differentiate in a sustainable way. This is clear in this example of the *Mabo* litigation and

its progressively greater 'land-ness' of subsequent cases in the litigation. The increasing dominance of 'land' and decreasing dominance of 'constitutional' matters is rendered quite clear by use of the visualisation.

Whilst this particular analysis and visualisation provides some clear relationship with the existing legal knowledge regarding *Mabo*, the same cannot be said so easily in relation to the rise and potential fall of proximity (Figure 5).[91]

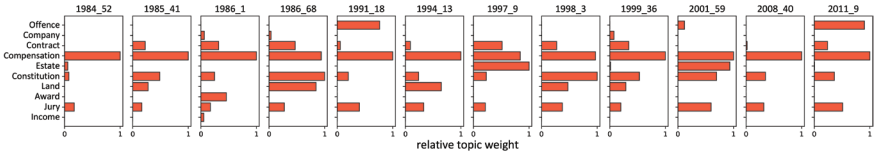

Figure 5: Topics for the 'proximity cases'

The cases used here are an attempt to trace a line of legal precedent and reasoning through key High Court cases that deal with the concept/definition of 'proximity' as it applies within Australian tort law. Like our approach to the *Mabo* litigation above, we have here visualised the topic modelling of a set cases identified *a priori* by the existing literature on proximity in aid of an integrative use of topic modelling with and alongside accepted legal interpretative methods. This approach benefits from an existing jurisprudence and scholarship regarding proximity which (and potentially against which) the computational results may be read.[92]

An alternative approach could be to use the use of the word/term 'proximity' within tort cases as a marker of its presence, and thus visualise those cases that met this criterion. However, the use of the term 'proximity' may unintentionally skew analysis, particularly if, as Andrew Robertson argues, the term 'proximity' itself has may have fallen away, but there is 'extensive continued use of the *substance* of the proximity concept in the High Court'.[93] Because of this challenge of identifying cases that engage with the test of proximity in tort law, we have

91 The cases that make up the 'proximity corpus' are drawn largely from the work of Andrew Robertson on proximity: see Robertson (n 46). The cases used are: *Jaensch v Coffey* (n 31); *Council of the Shire of Sutherland v Heyman* (1985) 157 CLR 424; *Stevens v Brodribb Sawmilling Co Pty Ltd* (1986) 160 CLR 16; *San Sebastian Pty Ltd v Minister Administering the Environmental Planning and Assessment Act 1979* (1986) 162 CLR 340; *Gala* (n 36); *Burnie Port* (n 36); *Hill v Van Erp* (1997) 188 CLR 159 ('*Hill v Van Erp*'); *Pyrenees* (n 44); *Perre v Apand Pty Ltd* (1999) 198 CLR 180; *Sullivan v Moody* (2001) 207 CLR 562 ('*Sullivan v Moody*'); *Imbree v McNeilly* (2008) 236 CLR 510; *Miller* (n 45).

92 Namely, in this article the work of Manderson and Robertson who have engaged in depth with proximity. This article is, unfortunately, not the venue for an extended engagement between the computational results and the oftentimes diverging interpretation of the place of proximity in Manderson and Robertson's work however. See Manderson, *Proximity, Levinas, and the Soul of Law* (n 38) 14; Manderson (ed), *Essays on Levinas and Law* (n 38); Manderson, 'Current Legal Maxims' (n 38); Robertson (n 46).

93 Robertson (n 46) 9 (emphasis added). See also at 21 where Robertson notes the 'metaphorical' character of proximity as used by the Court.

chosen instead the first 'integrative' approach.[94]

Here, with a larger number of key cases that mark the rise and fall of the test of proximity, there is no particular pattern or trend readily discernible other than that 'compensation' is the topic which unites the group of cases. This makes sense in so far as the cases concern the duty of care in tort (and thus are intimately related to the question of compensatory damages). However, we can discern a difference between earlier cases — for example *Jaensch v Coffey*, labelled here as '1984_52' which really established proximity within the Court, and later cases like *Sullivan v Moody*,[95] here labelled as '2001_59', which spelt the 'end' of proximity.[96] *Jaensch v Coffey* and cases surrounding it all are predominately — if not practically exclusively — about 'compensation'. The later *Sullivan v Moody* and other cases surrounding it in this genealogy of proximity cases have significant weights present in the topics of 'estate', 'constitution' and 'jury'. This is a clear point of differentiation between the earliest cases which established proximity and those afterwards which saw the Court move away and potentially reject the principle. It shows that proximity was developed, tested, accepted and potentially rejected or, if we follow Robertson, applied in a more limited way, within a series of cases that concerned a relatively wide range of secondary topics.

These observations regarding proximity, however, lack the more decisive learning gleaned from the integration of topic modelling in analysis of the *Mabo* litigation. In that case, there was a clearly discernible shift between the various *Mabo* cases, that accorded with and supported readings of the transition of those cases produced by traditional legal analysis. In relation to proximity, however, there seems to be no such decisive and clear learning that can be gleaned by integration of this method. This could be driven by the object of study in question. Proximity is a test or 'doctrinal object', that does not shift on a *topical* basis over time. It is, rather, a piece of doctrinal machinery that is applied within tort law, but to a variety of factual scenarios. The *Mabo* litigation and its development of native title jurisprudence does shift on a topical basis; from constitutional through to land-centred questions. This difference could be simply that there are different 'topical shapes' to each case study, and for proximity at least, this means that a chronologically read topic analysis does not add much to our existing analysis.

Recall too, that there is a dissenting voice in the work of Robertson regarding the 'end' of proximity. Robertson argues that proximity endures well after the otherwise

94 There is a significant project in developing and testing methods for identifying legal doctrine and its change over time by alternate means — including those that do not rely upon (or rely less upon) existing legal scholarship to identify cases for analysis. At this stage we believe it is possible only to do so by way of this integrative approach. Moreover, we believe also that this approach brings with it the practical and normative benefits outlined in the introduction to this article.

95 *Sullivan v Moody* (n 91).

96 Christian Witting, 'Tort Law, Policy and the High Court of Australia' (2007) 31(2) *Melbourne University Law Review* 569.

widespread view that its reign ended in 1999.[97] He demonstrates that it is applied in form — just not in name — in cases like *Graham Barclay Oysters Pty Ltd v Ryan*,[98] *Sydney Water Corporation v Turano*,[99] *Gifford v Strang Patrick Stevedoring Pty Ltd*,[100] and *King v Philcox*.[101] If Robertson's reading is correct, then seeking to see an 'end' or 'decline' of proximity's use by the Court is potentially misguided. So too would Robertson's view likely necessitate seeking an alternate approach from topic modelling as to computational method for analysis; if it is true that the test endures in practice, but that it is no longer named — and in fact as its use is denied and repudiated by the Court itself — then alternative approaches will have to be tested.

1 *Contextualising Topic Model Results in the Work of the Court*

While the results presented immediately above can be used to study the High Court of Australia and its judgments at various scales, it would be helpful to have a way of visualising relationships between cases in the greater context of the work of the Court, rather than with a view of particular cases or groups of cases in isolation as the previous two examples did. The challenge of doing so is that each of the more than 7,500 decisions of the High Court is represented by 10 topic weights. Mathematically speaking, each decision is represented by a point in a 10-dimensional space. In order to visualise the entire corpus, we need to project it on a 2- (or 3-) dimensional space, however we need to do so in such a way that local similarity is preserved, meaning that two cases that are similar in our 10-dimensional topic space, should ideally end up as two neighbours in the 2-dimensional representation. To this end we use a nonlinear embedding technique called Uniform Manifold Approximation and Projection ('UMAP')[102] to visualise the whole High Court corpus, to paint a topical portrait of the High Court as it were (Figure 6).

97 Robertson (n 46) 26–32.

98 (2002) 211 CLR 540.

99 (2009) 239 CLR 51.

100 (2003) 214 CLR 269.

101 (2015) 255 CLR 304.

102 Leland McInnes et al, 'UMAP: Uniform Manifold Approximation and Projection' (2018) 3(29) *Journal of Open Source Software* 861.

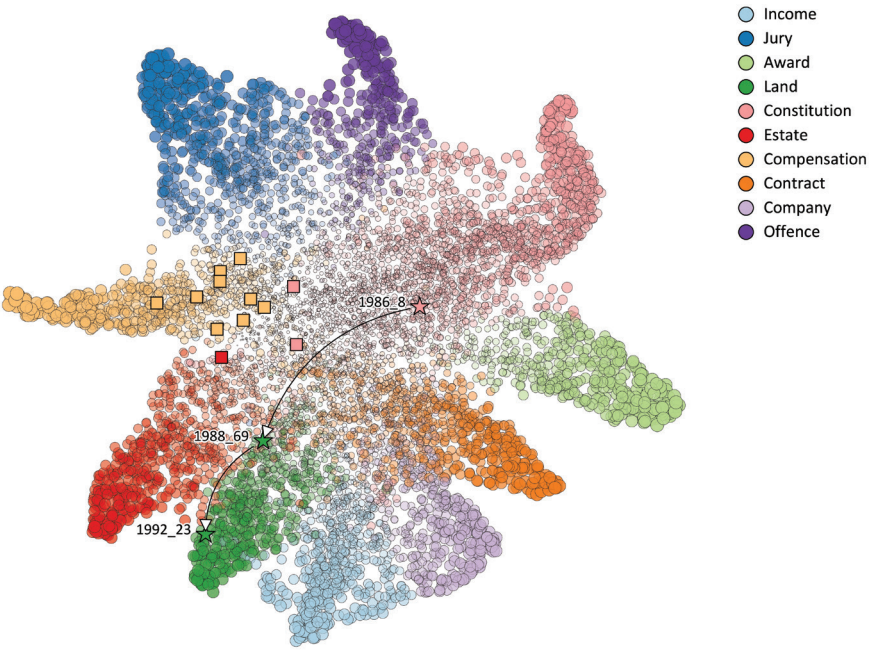

Figure 6: Embedding of the High Court of Australia corpus 1903–2017

Figure 6 displays a single point or circle for each judgment of the High Court in our corpus. The colour, size and location of each circle relative to others communicates further information about that judgment and its 'place' within the broader work of the Court; or more accurately, the place of the judgment within our own model of the Court's work. The Figure in its full-colour version displays ten distinct colours.[103] The colour of each circle denotes the dominant topic of the judgment — that is, the topic that dominates the judgment's classification according to our topic model. For example, purple data points are cases that have been classified as belonging to the topic labelled 'offence'. They are cases that our model has classified as relating to the criminal law, at least in large part. The size and opacity of each data point vary, reflecting the proportion of the topic allocation belonging to the dominant topic. For example, a smaller and more transparent circle denotes a case that is a mixture of several topics, with no strong or pronounced association with a particular topic. These are mostly found near the centre of the Figure. By contrast, a darker or less opaque and larger data point denotes a judgment that is more strongly or predominately associated with a single topic. Given this patterning, the larger and more densely coloured data points are located at the extremities of the 'tentacles'. That is, they are more 'fully within' the topic.

Within this visual representation we have plotted the key proximity cases by use

103 Rahmani (n 69).

of squares of a fixed size and opacity. The vast majority of these cases are located in the 'compensation region'. This corresponds with our analysis shown at Figure 5 above where the strongest or most dominant topic of the suite of proximity cases was the topic of 'compensation'. However, here we begin to see more clearly that there is some variation in the location of these proximity cases — and thus variation in their topical character. For example, the single red square corresponds to the case of *Hill v Van Erp*.[104] The classification of this proximity case as a matter of 'estate' seems anomalous. Proximity, after all, is a test primarily applied to the question of duty of care in tort. And yet, *Hill v Van Erp*, has been classified outside of the main body of tort cases (found in the 'compensation' area of the Figure), despite its 'leaning towards' the compensation case area in the Figure. Review of the case itself makes clear why this might be the case. The case concerned a question of the duty of care as between a solicitor and client, concerning beneficiaries to a will drafted by that solicitor, the process of executing a will and thus question of liability of a solicitor (and potentially other professional advisors) for resultant economic loss.[105] Proximity was engaged with in this case. In fact, the test of proximity was claimed by some to have been 'put to rest' by this case, and for others the case stands as the definitive turning-point away from proximity as it had been understood — it being the first duty case since Deane J, the most consistent advocate for 'proximity', had departed the bench.[106]

How might the location of the 'proximity' cases be interpreted in this Figure then? Taken as a group, the proximity cases for the most part never seem to make it 'deep' into the tort law territory (labelled in our topic model as 'compensation'). They seem at all times to be somehow 'peripheral' and by that peripheral location *differentiated* from the 'main body' of tort cases which make up the most opaquely coloured/most concentrated area of compensation/tort cases all found at the extremity of the 'compensation' tort law topic representation. Existing legal knowledge about the alleged 'rise and fall' of proximity has consistently understood its place as similarly peripheral, as never having been completely embraced by the Court. Moreover, the test — even during its rise — found its most significant application in cases that required adjudication or commentary regarding the existence of a duty of care in the first place. Perhaps the rarer nature of these cases before the Court itself drives the peripheral nature of the proximity cases' location when compared with the more straightforward questions of the quantum of damages or other more common issues

104 *Hill v Van Erp* (n 91). Shown as '1997_9' in Figure 5.

105 Des Butler, 'Proximity as a Determinant of Duty: The Nervous Shock Litmus Test' (1995) 21(2) *Monash University Law Review* 159; Butler, 'Managing Liability for Bystander Psychiatric Injury in a Post-*Hill v Van Erp* Environment' (n 36); Dan Fitzgerald, 'Showdown for Victoria: The High Court Judgment in *Van Erp v Hill*' (1997) 71(7) *Law Institute Journal* 33.

106 Manderson, *Proximity, Levinas, and the Soul of Law* (n 38) 132. But see Robertson (n 46) especially at 22.

of tort law that the Court deals with.[107]

We may also revisit the *Mabo* litigation cases which are plotted in Figure 6 as stars (for ease of visualising them their size and opacity are fixed). Arrows between these points have been inserted indicating the chronological progression of the cases. Note here how the progression of the cases begins within the constitutional law area, and then markedly shifts towards an ever 'deeper' or more pronounced 'land' character. Interestingly, the 1988 case — *Mabo [No 1]* — lies at the boundary between the land and compensation regions (in agreement with the results of Figure 9). The 'final' 1992 case — *Mabo [No 2]* — sits more deeply into the 'land' topic.

Given this, in a single Figure we are able to now visualise not only the relative change in topics between the cases, but also how they fit within the larger work of the court in a manner that accords with the doctrinal and legal-historical understanding of the evolution of the *Mabo* litigation. As Lisa Ford puts it in relation to the resolution of the *Mabo* litigation:

> In 1992, the High Court of Australia declared to the world that it would right the wrongs wrought by two centuries of settlement. This could all be done, the judges thought, by recognizing indigenous rights to land. They called these rights 'native title,' a title to land defined by native 'laws and customs' but subject to extinguishment, without compensation, by the sovereign. … [T]he court refused to inquire into the fiction of settler sovereignty — though the *Mabo* case, from the outset, had been mired in uncertainty about the relationship of territory, jurisdiction, and sovereignty in the Mer Islands in the Torres Strait.[108]

In this way, the case's position deeply within the 'land' topic and with movement away from areas like 'constitution' or state sovereignty provides some visual analogue to the argument put forward by Ford and that canvassed earlier by Fitzmaurice, that the *Mabo* litigation was characterised, interpreted and resolved as a matter of 'land law' rather than one relating to competing, overlapping or neighbouring claims of sovereignty per se.

B *Twenty-Topic Model*

These visual representations of the change and development of the *Mabo* litigation and of proximity are reinforced by review of the 20-topic model that was also fit

107 There is scope to test this question by a thorough and complete 'mapping' of all judgments/cases that apply or engage the test of proximity. The cases mapped here are those accepted as 'canonical' in the rise and fall of proximity. They are judgments that centre and focus upon the question of proximity, rather than those which might simply apply it. To pursue such analysis is likely more helpful in the lower courts — the District and Supreme Courts of each state and territory — where the test was applied at first instance, rather than in case of the High Court which engages with the question as a matter of law in an appellate fashion.

108 Lisa Ford, *Settler Sovereignty: Jurisdiction and Indigenous People in America and Australia, 1788–1836* (Harvard University Press, 2011) 207.

to the corpus. The 20-topic model brings with it the advantage of a greater level of granularity than the 10-topic model. And yet, the topics themselves seem still to 'fit' the established categories of law and the legal workload of the High Court.

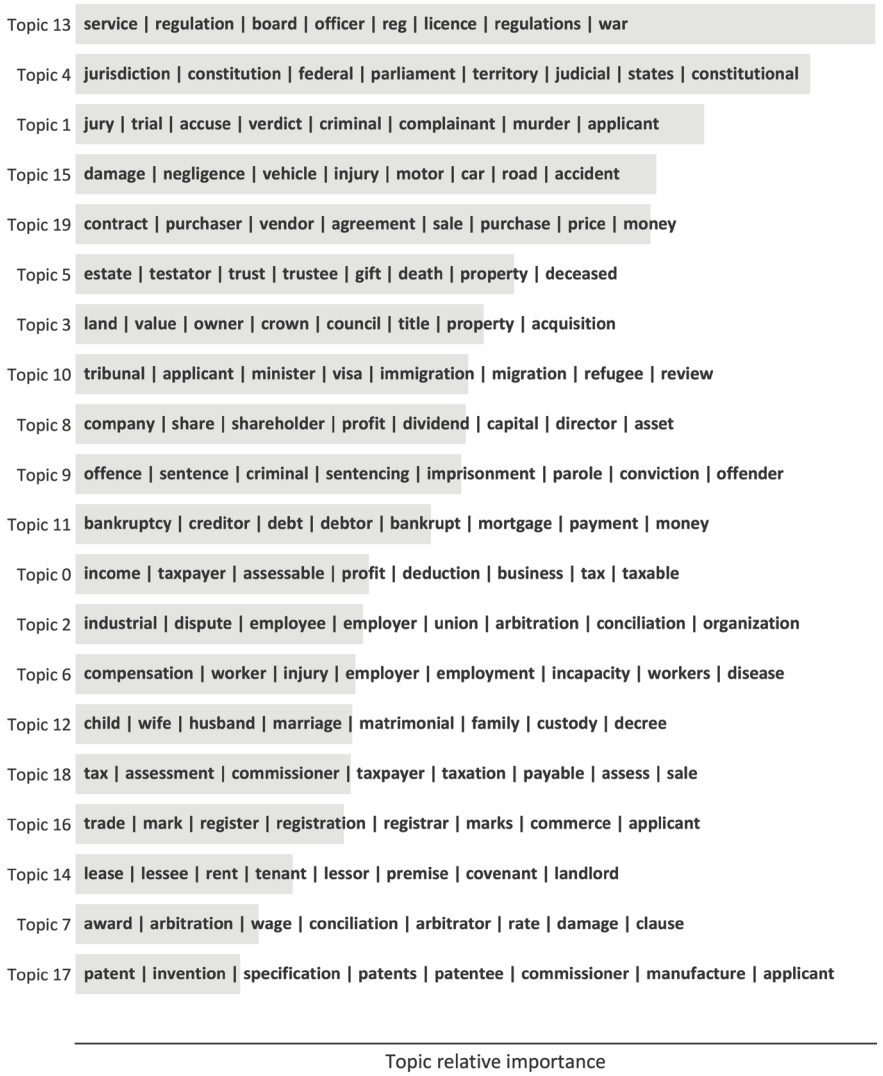

Figure 7: Twenty-Topic Model: Topics' relative importance

Whilst the 20-topic model (Figure 7) shares a range of topics that are similar to that of the 10-topic model (Figure 1), it seems also to 'explode' some of the topics generated in that earlier model. For example, the topic labelled 'compensation' is present in both topic models. However, in the 20-topic model, this is joined by conceptually and legally related topics like 'damage'. The 10-topic model 'compensation' included

reference to cases that involve injuries to workers, accidents, negligence and other material that would, in a standard taxonomy of legal materials, be separated from one another. This is achieved in the 20-topic model.

As to the question of proximity, Figure 8 displays the 20-topic model as applied to our group of proximity cases. All but one case is dominated by the 'damages' topic. The damages topic is one that focuses on tort liability, with the model listing the terms 'damage', 'negligence', 'vehicle', 'injury', 'motor', 'car', 'road', 'accident', 'loss', and 'care' as characteristic tokens or terms present in the cases that fit within this topic.

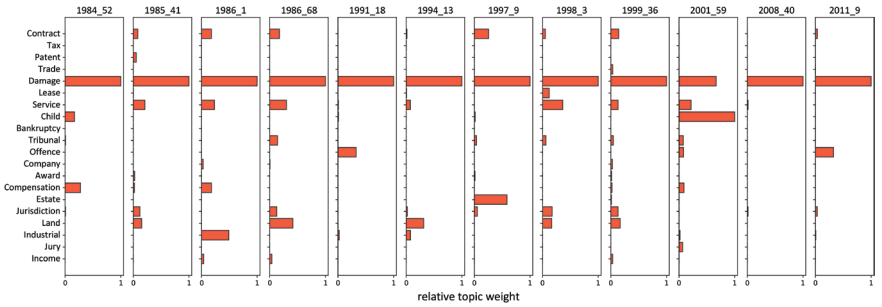

Figure 8: Twenty-Topic Model: Topics for Proximity Cases

Representing the proximity cases in our 2-dimensional visualisation reinforces this more granular reading of the proximity cases provided by the 20-topic model. In Figure 9, the proximity cases are again represented by squares of uniform opacity. They cluster now within the 'damage' topic area, with many of the cases reaching deep into that topic; a placement that demonstrates the centrality of 'damage' to those cases, where above the cases seemed potentially peripheral to the topic 'compensation'.[109]

The 2-dimensional representation of the 20-topic model also facilitates a more detailed view of the *Mabo* litigation cases, and of the arrangement of them across their most dominant topics. Represented again by stars, the litigation transitions in a manner that remains in accordance with their current legal interpretation: from a question of 'jurisdiction' (this time around) to a question more progressively of 'land' as the litigation proceeded. That is, a transition away from a case that we know to be factually concerned with the exercise of sovereign colonial power to extinguish native title by the State of Queensland towards resolution as a matter not of sovereign power/sovereignty/jurisdiction, but one of land law.

109 See Figure 6 and associated commentary.

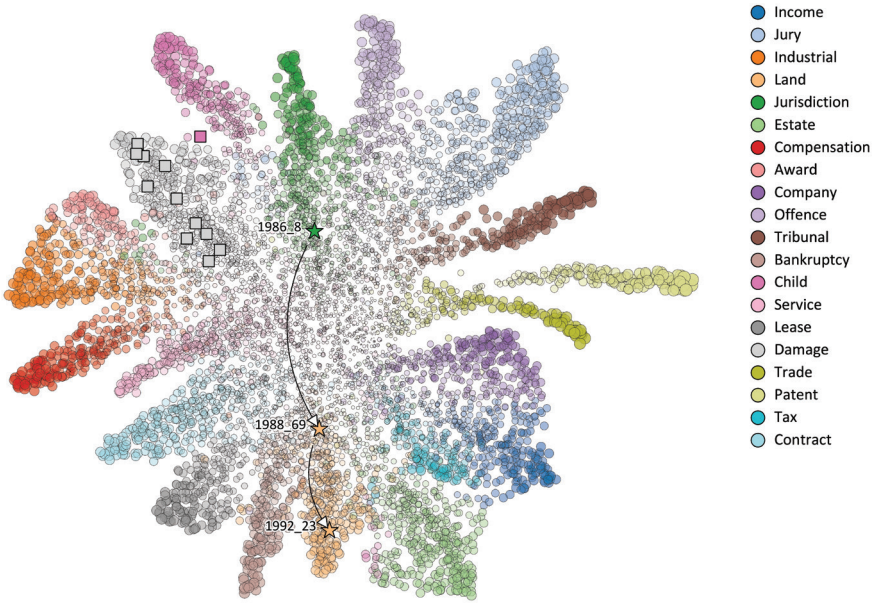

Figure 9: Twenty-Topic Model: Proximity (squares) and *Mabo* litigation (stars)

VI CONCLUSION

In this article we have reviewed the results of our pursuit of a 'finer-grained' mode of reading the law with the aid of computational analysis than has been generally engaged in before. We have demonstrated the use of topic modelling and associated visualisation to measure the emergence and development of particular legal doctrines through the rise and fall of 'proximity' as a test of foreseeability and the duty of care in tort and of the recent history of the High Court of Australia using the case studies of the *Mabo* litigation.

The results presented here demonstrate the ability for topic modelling techniques to engage with legal text at a level of granularity not generally achieved by other studies. From a technical or methods point of view, we have demonstrated the application of topic modelling to the combined judgments of the High Court of Australia that is useful for legal doctrinal research and interpretation. Moreover, this is an approach that brings with it some significant practical benefits for modelling a textual corpus of this size and type. Further, we believe that the approach used here also demonstrates one method — and more importantly the potential — of computer-aided quantitative analysis of legal text to both assist in the reading and analysis of law at varying distances, and to assist in the task of tracing the emergence and change of particular doctrines, to tell us something

of their character and to guide further analysis by other means, and to map and understand something of the history and development of a group of related matters before a particular court. These are all classic, and fundamental tasks of legal analysis — both practical and scholarly — and these techniques present an aid to the achievement of these tasks by these practical and theoretical advances.

There are two key jurisprudential and theoretical implications that arise from this work. First, it is notable that topic modelling is a technique that can successfully engage with legal text at a variety of greater and lesser levels of granularity. Second, and perhaps most significantly, the approach here outlines a way forward for the application of topic modelling — in combination with other data — to the task of tracing of the emergence and evolution of legal doctrines and of particular lines of case law both across time and within the context of the broader work of the High Court. Such shifting in scale and granularity of analysis achieved here by way of topic modelling is a signal feature of legal textual engagement and interpretation in contexts outside of computational approaches. As we have argued before:

> [L]egal interpretation and scholarship *require and emerge from* the explicit and ongoing shift between attention to sentence, vocabulary, word order and style which exemplifies the 'close' reading scale, and the more 'distant' reading scale marked by techniques of reconfiguring large collections of cases by topic, drawing out and constructing legal themes or subject matter, alongside the reading of legal texts at the very great distance of subsequent social practice and effect.[110]

The distant reading scale achieved above by visualising the relative location of both the *Mabo* litigation's key cases, and those associated with 'proximity' against the background of the Court's broader work tells us something about the character and focus of these cases; namely, how they 'fit' and how they also do not 'fit' into the particular areas of law to which they have been generally ascribed. So too do these visualised results indicate something fruitful regarding the application of this method and its potential to advance legal scholarship. The relatively peripheral location of the proximity cases to their 'home turf' of tort law ('compensation' in our 10-topic model) indicates, we hypothesise, something of their peripheral status within the Court's jurisprudence of tort law and the place of proximity within it. So too is the transition from a question of 'constitution' (in the 10-topic model) or 'jurisdiction' (in the 20-topic model) made by the key cases within the *Mabo* litigation reflective of the transition both in the factual and legal character of those cases.

Here we have used the *Mabo* litigation and 'proximity cases' to illustrate how computational approaches may be used to support and, to some extent quantify

110 Carter, Brown and Rahmani (n 3) 1306 (emphasis in original).

legal intuition and knowledge. There remains significant work to be done regarding both the technical and legal aspects of such computational analysis. For one, our use of the *Mabo* litigation and proximity line of precedent and reasoning reveals a significant limitation or challenge in that the corpus of cases for both areas of analysis were known *a priori*. Given this, one significant area for work must be to discover whether computational methods can assist legal scholars and practitioners to *discover* new connections between legal documents — and whether such computationally generated connections might be legally valid and useful. Another challenge, of a more fundamental nature, is work to name, develop and improve the methodological challenges that arise at the meeting of the computational and legal disciplines. This is an exciting and no doubt difficult task. However, it seems from this exercise at least, this is an essential task.