A Network Analysis of Dutch Regulations
Using the MetaLex Document Server

Rinke Hoekstra[1,2]

[1] Department of Computer Science / The Network Institute
VU University Amsterdam
rinke.hoekstra@vu.nl,
[2] Leibniz Center for Law, Faculty of Law, University of Amsterdam
hoekstra@uva.nl

Abstract. In this paper we explore the possibilities of using the Linked Data representation of all Dutch regulations stored in the MetaLex Document Server for the purposes of network analysis over the citation graph between regulations, both at the document level, and at the article level. We show that this is possible using relatively straightforward SPARQL queries, and present preliminary results of the analysis.

Keywords: Network Analysis, Linked Data, CEN MetaLex, Legislation, Regulation

1 Introduction

The MetaLex Document Server [2][3] (MDS) hosts all Dutch legislation and treaties available from spring 2011 onwards. The server was developed to overcome the limitations of the publicly available legislation published through the Wetten.nl portal of the Dutch government. Regulations available on MDS are published in two formats, CEN MetaLex, a flexible and jurisdiction agnostic XML format for publishing legal sources, and as Linked Data in RDF format. The Linked Data is browsable through a Pubby[4] interface that operates on top of a 4Store SPARQL endpoint. The Linked Data format is highly suitable for graph analysis as the format itself (RDF) forms a graph of interconnected Web resources (URIs). At the time of this analysis (April 22, 2013), the MDS hosted 280,394,1322 triples across 33,643 document versions.

In this paper, we show how we can run tailored network analyses of the Dutch regulations published in MDS using relatively straightforward SPARQL queries. We present preliminary results of these analyses, and discuss the benefits of using Linked Data as source for network analysis. The networks and analyses discussed in this paper are published separately on Figshare.com [3].

[3] See also http://doc.metalex.eu
[4] See http://github.com/cygri/pubby.

Measure	Document	Article	Factor
Number of nodes	14935	64018	4.286
Number of edges	33819	80082	2.368
Average degree	2.264	1.251	0.553
Avg. Weighted degree	9.117	3.749	0.411
Network diameter	16	8	0.5
Average path length	5.479	1.316	0.240
Avg. Clustering Coefficient	0.09	0.0021	0.023
Connected Components	492	7262	14.76
Number of SCC's	14019	63303	4.516

Table 1. Network properties

1.1 Regulations in the MDS

Regulations in the MDS are represented using the CEN MetaLex ontology, an automatically generated ontology of the Basiswettenbestand (BWB, the database underlying wetten.nl), the OPMV provenance vocabulary, the W3C Time Ontology, and the Simple Event Model (SEM). Of most importance to us here is the CEN MetaLex ontology. It provides vocabulary for distinguishing levels of description of regulations along the FRBR levels of *work*, *expression* and *manifestation*. In the MDS, every regulation is described both at the work level (e.g. 'the Income Tax Law') and at the expression level (e.g. 'the Income Tax Law of January 1st, 2013'). Every expression level resource is linked to its work via a `metalex:realizes` property.

Expression-level citations allow analysis of the citation network of regulations *through time*. Unfortunately, the targets of citations from regulations are not explicitly linked to a specific version, but only to the work identifier. This means that the resulting citation graph would be hugely disconnected: many nodes (the expressions) have only outgoing links, while other nodes (the works) have only incoming links. Consequently, measures such as betweenness centrality, clustering coefficient, connected components and network diameter will give very little information of the connections between regulations at the work-level. For this reason, the analysis presented in this paper only takes into account citations aggregated to the *work* level.

1. A citation from an expression level to a work, will be represented as a citation between works.
2. Two citations to separate expressions of a single work will only be counted once.
3. The highest level of detail of a citation is the *article* level, i.e. the most specific, uniquely and independently citable part of a regulation.

Table 2. Top-10 Betweenness Centrality

Rank	Name	Value
1	Algemene wet bestuursrecht	7007741
2	Wet milieubeheer	2172441
3	Besluit omgevingsrecht	1667495
4	Besluit algemene regels voor inrichtingen milieubeheer	948497
5	Wet op de economische delicten	770968
6	Wetboek van Burgerlijke Rechtsvordering	696456
7	Wet op het hoger onderwijs en wetenschappelijk onderzoek	687873
8	Wet op het financieel toezicht	664934
9	Algemene douanewet	616671
10	Circulaire bodemsanering 2009	561465

Table 3. Top-10 PageRank

Rank	Name	Value
1	Algemene wet bestuursrecht	0.0152
2	Wetboek van Burgerlijke Rechtsvordering	0.0117
3	Wet gemeenschappelijke regelingen”	0.00803
4	Wet openbaarheid van bestuur	0.00785
5	Wetboek van Strafvordering	0.00723
6	Grondwet	0.00712
7	Algemene termijnenwet	0.00668
8	Wet structuur uitvoeringsorganisatie werk en inkomen	0.00638
9	Kaderwet zelfstandige bestuursorganen	0.00623
10	Vreemdelingenwet 2000	0.00597

Table 4. Top-10 Indegree

Rank	Name	Value
1	Algemene wet bestuursrecht	426
2	Bezoldigingsbesluit Burgerlijke Rijksambtenaren 1984	336
3	Archiefwet 1995	278
4	Werkloosheidswet	265
5	Wet op de arbeidsongeschiktheidsverzekering	236
6	Ziektewet	220
7	Warenwet	210
8	Algemene Wet Bijzondere Ziektekosten	207
9	Wet op het voortgezet onderwijs	204
10	Zorgverzekeringswet	119

2 The Level of Regulations

In order to answer the question "What is the most important or influential regulation in the Netherlands?" we can analyse the network of co-citation between regulations as found in the MDS. Since all elements of a regulation (read articles, chapters, paragraphs etc.) are represented in RDF as part of a named graph, we can build this network by running a SPARQL query that simply returns the graph URIs of the source and target of every citation, respectively $?s$ and $?t$:[5]

```
PREFIX metalex: <http://www.metalex.eu/schema/1.0#>
PREFIX bwb: <http://doc.metalex.eu/bwb/ontology/>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT DISTINCT ?s ?s_title ?t ?t_title WHERE {
GRAPH ?s {
     ?s_ref metalex:cites ?t_id .
   }
   GRAPH ?t {
     ?t_id a ?type .
   }
   OPTIONAL {?s dcterms:title ?s_title }.
   OPTIONAL {?t dcterms:title ?t_title }.
}
```

For readability purposes, we also retrieve the titles of both regulations ($?s_title$ and $?t_title$). The result is stored in two separate CSV files, one for the edges (with columns "Source" and, "Target"), and one for the nodes ("Id", "Label"), and loaded as a graph in Gephi.[6] Table 1 shows the network properties of the resulting citation graph. Tables 2, 3 and 4 show the top ranking nodes (documents) for Betweenness, PageRank and Indegree, respectively. The resulting graph is depicted in Figure 1, where nodes and edges are colored according to the applicable module, and size of nodes corresponds to the PageRank score.

Betweenness centrality measures the relative number of shortest paths that run through a node. The intuition is that nodes with a high betweenness centrality are important for connecting separate parts of a graph. In other words, documents with a high betweenness centrality connect different, otherwise unconnected parts of the Dutch regulations. The "Algemene Wet Bestuursrecht" (AWB), the general administrative law is high in the list as it is large, and touches upon virtually all regulations that concern the Dutch government. Environmental ('milieu' and 'omgeving'), economic, and civil ('burgerlijk') laws have similar qualities. The 'circulaire' is a type of regulation that has the specific function of bringing together aspects of multiple regulations for a specific target audience.

[5] Note that this query will not return all citations, but only one per source/target pair. Also, the endpoint at http://doc.metalex.eu is limited for performance reasons, so it may not return the same results as used for this analysis.

[6] See http://gephi.org.

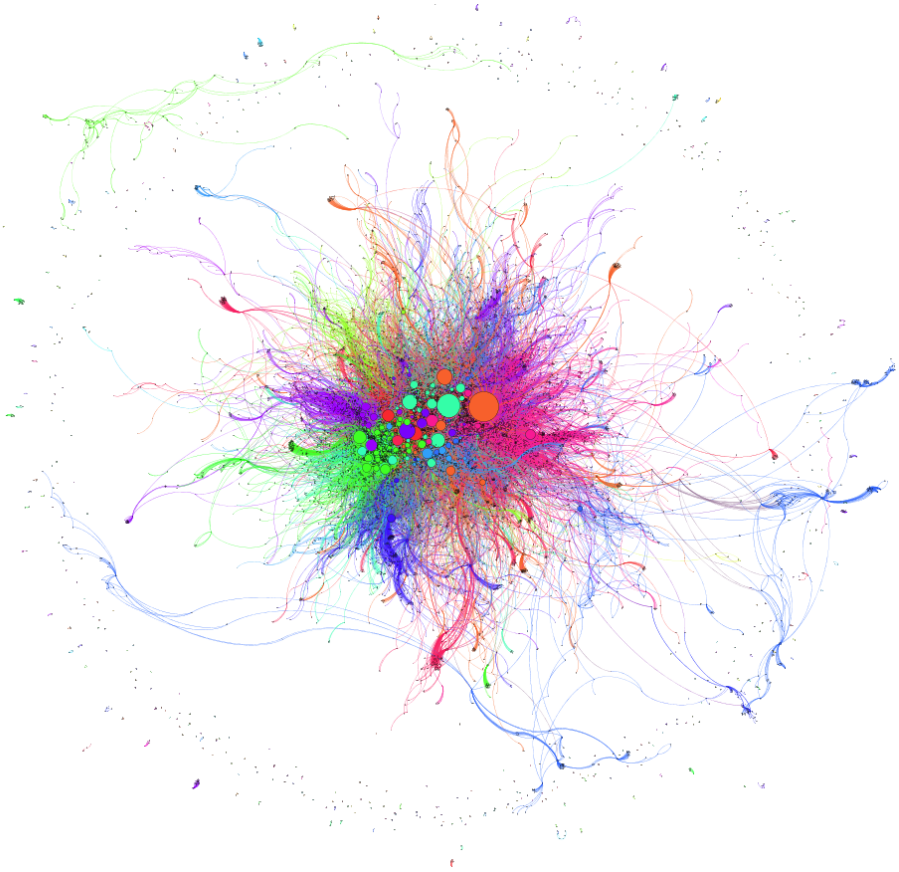

Fig. 1. Citation network between regulations at *work* level.

Indegree measures the number of incoming edges to a node. This is an absolute measure of importance. Again the AWB tops the list, but other regulations reflect on penal procedures ('strafvordering'), unemployment ('werkloosheid'), healthcare ('ziekte'), education ('onderwijs') and salaries ('bezoldiging') of civil servants at the national level. PageRank [4], one of the algorithms used by Google, is a relative measure of importance. Again the AWB tops the list, but the constitution ('Grondwet') makes a first appearance, as well as regulations concerning immigrants ('vreemdelingen'), freedom of information ('openbaarheid').

2.1 Evaluation

We compared the various network measures to a list of "important" regulations, i.e. those listed on Wikipedia[7] as belonging to the category of Dutch law. This is

[7] See http://nl.wikipedia.org/wiki/Categorie:Nederlandse_wet.

Table 5. Comparision to laws listed on Wikipedia

Measure	**Recall** (0.5x)	**Recall** (1x)	**Recall** (2x)	**Precision** (2x)	**F-Score** (2x)
PageRank	0.311	0.444	0.617	0.309	0.411
Indegree	0.296	0.474	0.612	0.306	0.408
Degree	0.260	0.423	0.551	0.276	0.367
Betweenness	0.240	0.388	0.536	0.268	0.357

a very imprecise measure, but it indicates at least a basic notion of importance. We normalized all names to lower case, and removed those laws from the target set Wikipedia that do not have a direct match with any regulation in our set. The reason is that many of the regulations listed on Wikipedia are listed by citation title, rather than the full title. This reduced the list from 315 regulations initially, to 196 regulations (this is more than the 180 we could obtain by querying the wetten.nl portal directly). It should be noted that it is relatively straightforward to improve this number, e.g. by retrieving citation titles from the MetaLex Document Server, and using a simple edit distance or bag of words comparison between titles.

Table 5 shows recall and precision for PageRank, betweenness centrality, degree and in degree as they apply for varying sizes of the result set. Precision only applies in cases where the result set is larger than the target set (i.e. the length of the list of regulations from Wikipedia).

The results for this comparison shows that PageRank and in degree compete for the first place with respect to the ability to predict occurrence of a regulation in the Wikipedia category. However, PageRank performs consistently better over multiple result set sizes. Only in the case where the result set size matches the target set size exactly (Recall 1x), the in degree measure results in higher recall. It should be noted that only with a result set of 61 times the target set size, recall for PageRank and in degree reaches 100%: we need to consider approximately 80 percent of all regulations in our graph. For degree this point lies at around 50 times the target set size.

3 The Level of Articles

If we consider citations to- and from the *article* level, i.e. we look for an answer of the question "What is the most important or influential article in the Netherlands", we design a SPARQL query that does not aggregate to the graph level, but considers the citing article itself:

```
PREFIX metalex: <http://www.metalex.eu/schema/1.0#>
PREFIX bwb: <http://doc.metalex.eu/bwb/ontology/>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT DISTINCT ?s_ref ?s_title ?t_id ?t_title WHERE {
```

Table 6. Top-10 Betweenness Centrality

Rank	Name	Value
1	Wet op de omzetbelasting 1968, Bijlage I	829.5
2	Wijzigingswet Wet luchtvaart (Regelgeving burgerluchthavens en militaire luchthavens), Artikel X	504
3	Warenwet, Artikel 1	492.5
4	Warenwet, Artikel 3	436.5
5	Wet vergoedingen adviescolleges en commissies, Artikel 2	423
6	Pensioenwet BES, Artikel 1	373
7	Administratiebesluit Bijzondere Ziektekostenverzekering, Artikel 1	362
8	Besluit inbeslaggenomen voorwerpen, Artikel 1	319
9	Rijkswet wijziging Statuut in verband met de opheffing van de Nederlandse Antillen, Artikel I	306
10	Wet openbaarmaking uit publieke middelen gefinancierde topinkomens, Artikel 2	294

Table 7. Top-10 PageRank

Rank	Name	Value
1	Algemene wet bestuursrecht"	0.00262
2	Archiefwet 1995	0.00242
3	Wet op het financieel toezicht	0.00196
4	Zorgverzekeringswet	0.00175
5	Algemene Wet Bijzondere Ziektekosten	0.00167
6	Bezoldigingsbesluit Burgerlijke Rijksambtenaren 1984, Bijlage B	0.00162
7	Wet op het voortgezet onderwijs	0.00159
8	Wet bescherming persoonsgegevens	0.00150
9	Wet op de omzetbelasting 1968	0.00148
10	Werkloosheidswet	0.00147

Table 8. Top-10 Indegree

Rank	Name	Value
1	Algemene wet bestuursrecht	558
2	Werkloosheidswet	453
3	Wet op de arbeidsongeschiktheidsverzekering	453
4	Ziektewet	493
5	Archiefwet 1995	398
6	Wet op het voortgezet onderwijs	364
7	Wet op het financieel toezicht	361
8	Algemene Wet Bijzondere Ziektekosten	342
9	Wet werk in inkomen naar arbeidsvermogen	327
10	Zorgverzekeringswet	326

```
    GRAPH ?s {
     ?s_ref metalex:cites ?t_id .
   }
   GRAPH ?t {
     ?t_id a ?type .
   }
   OPTIONAL {?s dcterms:title ?s_title }.
    OPTIONAL {?t dcterms:title ?t_title }.
}
```

As one can see, the query is very similar to the one we use for the document level citations, but instead of $?s$ and $?t$, we look for $?s_ref$, the identifier of the CEN MetaLex element that cites, and $?t_id$, the identifier of the cited resource. Citations in CEN MetaLex are represented *inline*, that is, the RDF representation does not contain explicit `metalex:cites` predicates on e.g. articles or members. The citations originate from resources at a lower level in the `metalex:partOf` hierarchy. Unfortunately the triple store of MDS (4Store) does not support SPARQL 1.1 property paths[8], and ascending the `metalex:partOf` hierarchy via unions in the SPARQL query is very expensive (read: slow).

We therefore reconstruct the article identifier from the citation-level identifier by parsing the transparent URI of the citing element. For instance, the URI:

`http://doc.metalex.eu/id/BWBR0002634/hoofdstuk/XII/artikel/33/lid/2/al/2/extref/1/nl/2012-01-01`

is used to construct:

`http://doc.metalex.eu/id/BWBR0002634/hoofdstuk/XII/artikel/33/nl/2012-01-01`

The next step is to reconstruct the work-level identifier for the article, by removing any language tag or timestamp information:

`http://doc.metalex.eu/id/BWBR0002634/hoofdstuk/XII/artikel/33`

Note that we could also have retrieved the work-level identifier directly through the SPARQL query, if we had queried along the `metalex:realizes` predicate. However, this would have introduced yet another expensive join in the query. Table 1 shows details of the resulting citation graph, and Figure 2 depicts a rendering of the graph where nodes are sized according to PageRank, and colored according to module.

Table 10 shows the values for *in degree* of the Algemene wet bestuursrecht (AWB, administrative law) per chapter. The law itself is cited a total of 558 times, where individual parts of the law are together cited 207 times:[9] 37% of all citations are to *parts* of the law.

[8] See http://www.w3.org/TR/sparql11-property-paths/

[9] Note that the document level statistics for indegree, show an aggregated number. It only counts multiple citations between two laws once, where the article level statistics count every citation.

Table 9. Power law

Measure	Article
PageRank	0.6996583
Degree	2.896214
Indegree	2.19982
Betweenness	3.658579

Table 10. Indegree per part of the Algemene wet bestuursrecht (AWB)

Part	Indegree
Algemene wet bestuursrecht	558
Chapter 9	52
Chapter 6	48
Chapter 8	37
Chapter 7	33
Chapter 10	9
Chapter 5	9
Appendix 2	5
Chapter 3	4
Chapter 4	4
Chapter 2	3
Appendix 3	3

4 Is the Law like the Web?

Both the law and the Web are man-made networks of interlinked documents. It is a valid question to ask whether the graph properties of both networks resemble eachother. Or, put in another way, is the distribution of information within the body of Dutch regulations specific for the domain of law, or is it similar to the more organically grown body of information that resulted in the anarchy of the Web?

To begin to answer this question, we can consider two important properties of the Web: it is *scale free* (an *ultra* small world), and it contains a single *giant strongly connected component* (SCC) that contains roughly a third of all pages [1]. Scale free networks have a degree distribution that follows a power law. Table 9 lists the result of fitting various distributions to a power law function, using the *igraph* package in R.[10] This suggests that indeed the degree distribution follows a power law, where $\alpha = 2.19982$, and the citation graph of Dutch legislation is scale free.

The structure of the Web resembles a bowtie (Figure 3, [1]) with at its heart a giant SCC, and incoming and outgoing nodes on the left and right, respectively. There are several smaller components that are wholly unconnected to the giant SCC, as well as tubes, that bypass the SCC, and tendrils that originate from the

[10] See http://igraph.sourceforge.net/.

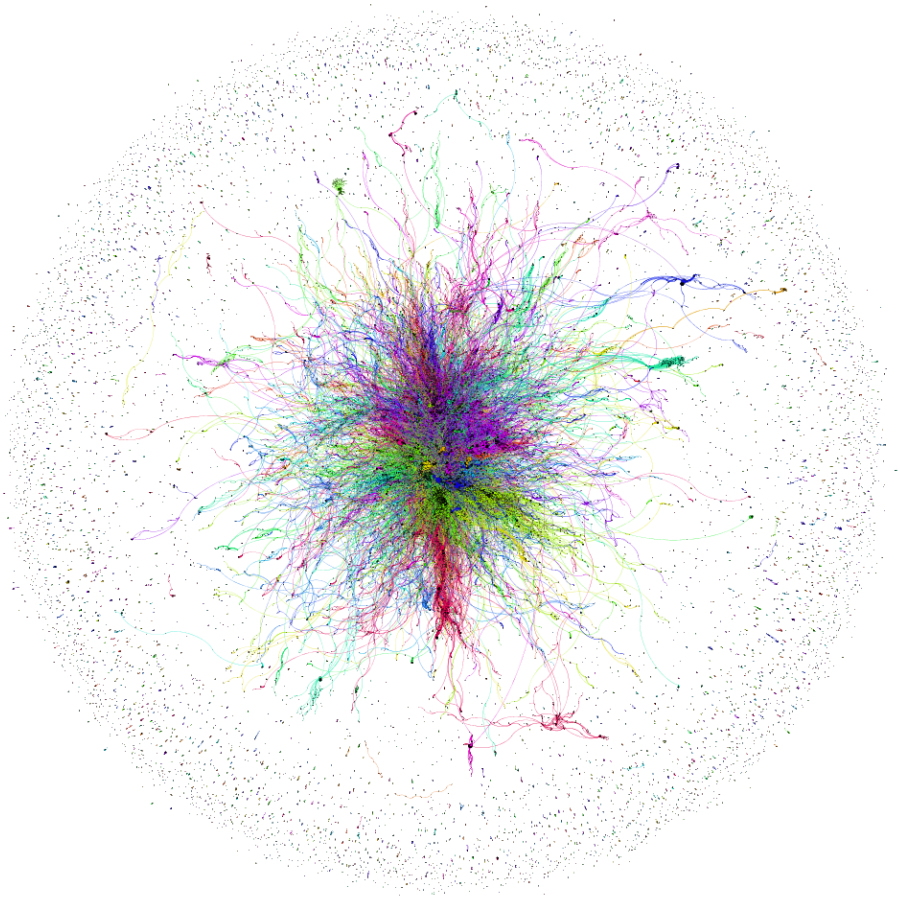

Fig. 2. Citation network between articles at *work* level.

incoming and outgoing nodes. The giant SCC covers approximately one quarter of all Web pages.

The *document*-level network of Dutch regulations contains 14019 SCCs, one of which is 816 nodes in size, where 74 others have between 2 and 6 nodes: the vast majority are single-node SCCs. This means that although we do have a similar situation with a giant SCC, the network as a whole is not as connected as the Web. For the *article*-level network, the situation is even more different: 63303 SCCs with no giant SCC (maximum of 12 nodes) and 501 SCCs of size larger than one. This can be explained by the decreased likelihood of 'random' edges between nodes in a curated network in general, and the relatively smaller chance of an edge between articles than when edges are aggregated to document level. Indeed, Table 1 shows a much lower average degree for the article-level network. The average path length and network diameter are much smaller for

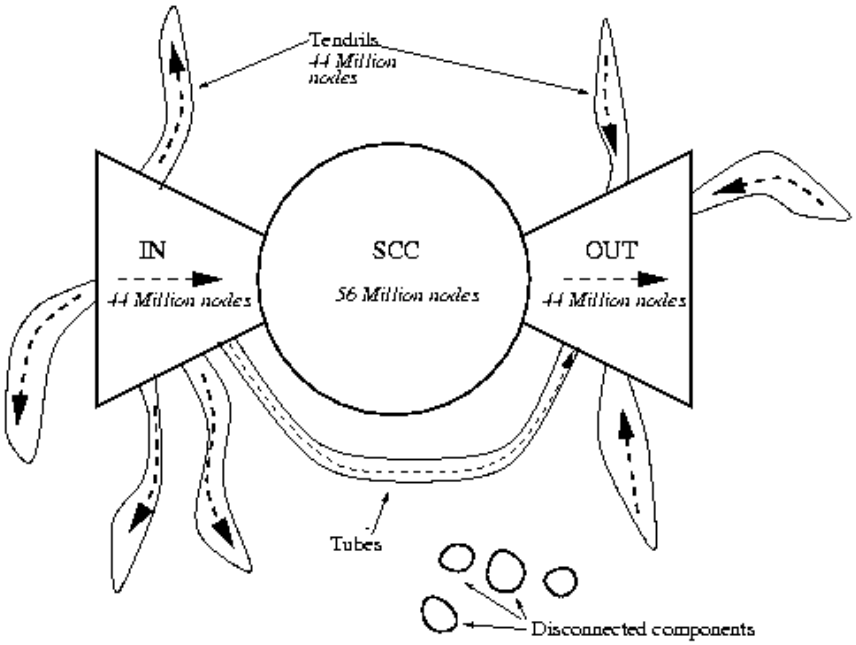

Fig. 3. The bowtie structure of the web, according to [1].

the article-level network as well, this is likely due to the much larger number of components.

5 Discussion

In the preceding we presented preliminary results of a network analysis of Dutch regulations stored in the MetaLex Document Server [2]. The networks were constructed using straightforward SPARQL queries against the MDS endpoint, requiring only minimal transformation to analyzable form. The analysis itself was performed using a variety of off-the shelf tools, primarily R and Gephi.[11]

Because of the preliminary nature of this experiment, it is hard to draw any conclusions with respect to what the analysis tells us about Law. We compared the selection of regulations based on various network metrics to a sample list of regulations from the Dutch Wikipedia page. This is a relatively arbitrary selection, and the results are correspondingly in-definitive. However it does point in an interesting direction: do network metrics on citations between regulations tell us anything about the *importance* or *role* of those regulations? Also, it

[11] For R, see http://www.r-project.org.

would be interesting to see whether citations to articles indicate e.g. a high representation of *definitions* in the cited article [5].

Secondly, we compared network properties of the document- and article level networks to that of the Web, and concluded that both networks are scale free, but the connectedness of regulations is much lower than that of the Web. The lower connectedness makes it easier to distinguish *modules* in the set of regulations. It would be very interesting to see how the results of generic module recognition algorithms correspond to actual topics in legislation.

Acknowledgments

This publication was supported by the Dutch national program COMMIT.

References

1. Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 309–320, Amsterdam, The Netherlands, The Netherlands, 2000. North-Holland Publishing Co.
2. Rinke Hoekstra. The MetaLex Document Server - Legal Documents as Versioned Linked Data. In Harith Alani and Jamie Tailor, editors, *Proceedings of the 10th International Semantic Web Conference (ISWC 2011)*, page 16. Springer, 2011.
3. Rinke Hoekstra. A Network Analysis of Dutch Regulations. Figshare.com, 2013. doi:10.6084/m9.figshare.689880.
4. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.
5. Radboud Winkels and Rinke Hoekstra. Automatic Extraction of Legal Concepts and Definitions. In Burkhard Schäfer, editor, *gal Knowledge and Information Systems, Jurix 2012: the Twenty-Fifth Annual International Conference*, pages 157–166, Amsterdam, 2012. IOS Press.