TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Raigo Aljand

On the Topological Structure of Legal Act Networks: Empirical Evidence from Legal Acts Networks of UK and Estonia

Master's thesis

Supervisor: Innar Liiv
Ph.D.

Tallinn 2019

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Raigo Aljand

Õigusakti võrkude topoloogilisest struktuurist: empiirilised tõendid Suurbritannia ja Eesti seadusaktide võrgustikest

magistritöö

Juhendaja: Innar Liiv
Ph.D.

Tallinn 2019

Author's declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Raigo Aljand

January 6, 2019

Abstract

I analyze the structure of law in the countries of the United Kingdom and Estonia. I study the topological structure of legal acts. In the network of a legal act, a vertex is a section of the act and an edge is a directed reference from a section of that legal act to another section of the same legal act. I observe the topological structure using visualization and statistical network measures. I look at the degrees of vertices and study the degree distribution. I calculate the clustering coefficient for the graph and its vertices. I also calculate the average shortest path, diameter and radius of the graph. I observe the robustness of the network by removing nodes from the network randomly or by targeting important nodes. I observe the degree correlation of the graph. I conclude that legal acts are scale-free graphs like many other complex system networks. As an additional interdisciplinary experiment, I compare the topological structure of the UK law with the measurements of the impact of law by the World Bank Group in the Doing Business report.

This thesis is written in English and is 52 pages long, including 5 chapters, 23 figures and 6 tables.

Annotatsioon
Seaduse arvutusliku keerukuse hindamine graafiteooria ja tekstianalüütika abil

Analüüsin seaduse struktuuri Suurbritannia ja Eesti seadustes. Uurin seadusaktide topoloogilist struktuuri. Seadusakti graafis on üks tipp seaduseakti peatükk ja üks kaar on seadusakti-sisene viide peatükilt peatükile. Vaatlen võrkude topoloogilist struktuuri, kasutades visualiseerimist ja arvutades võrgu statistilisi mõõteid. Uurin tippude astmeid ja astmete jaotust. Järeldan, et astmete jaotus järgib power-law funktsiooni ja võrgud on skaalata (*scale-free*) struktuurid. Arvutan klastrikoefitsiendi (*clustering coefficient*) nii graafile kui ka tippudele. Leian keskmise lühima tee, graafi diameetri ja raadiuse. Järeldan, et Inglismaa seadusi ei saa lugeda väikse-maailma võrkudeks. Vaatlen graafi vastupidavust vigadele ja rünnakutele, eemaldades tippe, kas juhuslikult või tähtsamaid valides. Vastupidavuse analüüsis tuleb samuti välja, et tegemist on skaalata (*scale-free*) graafidega nagu ka muude valdkondade keerulised graafid. Uurin tippude naabrite astmeid. Arvutan Suurbritannia seadustele Andres Kütti keerukusmõõte. Leian, et Eesti seadused on märkimisväärselt pisemad kui Suurbritannia seadused ja ka natuke hierarhilisemad. Eraldiseisva interdistsiplinaarse eksperimendina võrdlen minu graafide topoloogilist struktuuri Maailmapanga tehtud mõõdetega seaduse mõjust nende aruandes Doing Business. Võtan seadused, mida kasutati Doing Business aruande loomiseks ja arvutan nende jaoks topoloogilise struktuuri mõõtmeid. Seejärel võrdlen ma neid mõõtmeid Doing Business edukuse koondarvuga, mille Maailmapank välja mõtles. Leian, et ei ilmne selget mustrit võrkude topoloogiliste struktuuride ja Doing Business aruande tulemuste vahel. Võrdlen ka seaduste võrkude topoloogilist struktuuri muude domeenide topoloogiliste struktuuridega.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 52 leheküljel, 5 peatükki, 23 joonist ja 6 tabelit.

Nomenclature

API	Application programming interface
DC	Detached Component
DSM	Design Structure Matrix [1]
GCC	Giant Connected Component or Giant Weakly Connected Component
GSCC	Giant strongly connected component
LZMA	Lempel–Ziv–Markov chain algorithm
P2P	Peer-to-peer
URL	Uniform Resource Locator
XML	Extensible Markup Language

Contents

List of Figures

List of Tables

1 Introduction

Networks are a research topic that has been well studied and applied into many complex systems [2–16]. They are a mathematical tool that allows to abstract the role of an actor to a simple model and allow the study of the complex system as a whole. While the study of a single actor is valuable, study of the system as a whole may reveal other interesting behavior. After simplifying a domain into a network, one can use visualization tools like drawing a graph, building an adjacency matrix or constructing a heatmap. Or one can use statistical methods to condense a behavior of the graph into a single number or easily visualizable diagram.

Looking at the degree distribution of a network, the network falls into one of two categories: homogeneous networks and heterogeneous networks. The degree of a node is the number of its neighbors.

Homogeneous networks are well simulated using completely random networks. A homogeneous network will usually have a few nodes with a very low degree, a few nodes with a very large degree and a lot of nodes having a somewhat average degree. On a degree distribution this will look like a Poisson or normal distribution. An example of this network can probably be seen in humans and their connections.

Heterogeneous networks will have a lot of nodes with low degrees and a few nodes with a very high degree. On a degree distribution, this network is characterized by a constantly decaying line. The degree distribution likely follows a power-law function and if so, is called a scale-free network. Many complex system networks are scale-free networks, for example the World Wide Web [9].

There are also other differences between homogeneous networks and heterogeneous networks that can be observed with statistical measures: robustness simulation behavior, betweenness, clustering coefficient and more.

This work is exploring the topological structure of complex previously unstudied networks. There has been other work in the domain of law, but with a different focus and data set [2–16]. I study the network of references between sections in the legal acts of the United Kingdom and Estonia. I also try to compare the topological structure of UK law with the impact of that law. The source code is online and available on GitHub.[1]. The main goals of this thesis are:

[1] `https://github.com/raigoinabox/tallinn-tech-legislation-parser`

- acquisition of the data, that is, examples of the UK and Estonian law;
- parsing of references in the UK and Estonian law;
- a thorough statistical analysis of the network of references;
- as an additional interdisciplinary experiment, a comparison of the topological structure of UK networks to the results of the Doing Business report.

In section 2, I describe work done by others. In section 3, I explain how I obtained the empirical data, give some simple information about the data and describe the methodologies used in this work. In section 4, I describe the topological structure of UK and Estonian laws. I also compare the topological structure of UK law with the results of the Doing Business report. In section 5, I summarize the results and propose possible further work.

2 Related work

I will first present research that has been done in other domains using topological structure analysis, followed by research done using networks in the domain of law.

2.1 Topological structure analysis

Bu et al [2] analyzed the topological structure of budding yeast. While previous research focused on individual proteins or protein pairs, they wished to study the interaction of proteins as a whole. They used spectral analysis and clustering coefficient to discover quasi-cliques and quasi-bipartites in the network. Then, using a database of known functions of proteins, they categorized the quasi-components. They were able to confirm the findings of others, identify proteins whose attributed functions may be incorrect and identify functions of new proteins that were not yet identified.

Ripeanu, Foster and Iamnitchi [3] studied the topological structure of the peer-to-peer Gnutella network. They built a crawler to discover the structure of the P2P network and analyzed it. From the topological structure they could make conclusions about network performance, reliability and scalability. They discovered that the Gnutella P2P structure does not match well with the underlying internet structure and made suggestions for improvement.

Fagiolo, Reyes and Schiavo [4] analyzed the World Trade Web and its evolution over time. The World Trade Web is the network of countries trading with each other. They applied weighted-network analysis to verify the results of other research on binary World Trade Web networks. They discovered, contrary to the results of previous studies, that most links are weak links and that countries with strong trade relationships are more clustered.

2.2 Legal act interconnection networks

Liiv, Vedeshin and Täks [17] studied the visualization of Estonian law. Similar to this work, they used references between sections to transform a legal act into a network. Natural text parsing was used to find these references and their location within the text. They used conformity analysis to cluster similar nodes together. Clustering similar nodes together made them easier to visualize and to visually notice patterns within the network.

Bourcier and Mazzega [18] studied codified French law. They had gained access to the codified French Environmental Code (LEC). LEC was divided into an eight-layered hierarchy: parts, books, titles, chapters, sections, subsections, paragraphs and articles, where each lower-layer object is a part of the higher-layer object. They used text parsing to discover the topological structure of LEC: they parsed out every type of object, its location and references to articles. They discovered that the total number of each type of object wasn't monotonously increasing, most articles were relatively simple and there were a few complex articles where complexity seemed to gather.

Instead of using references to create a graph, Täks et al [19] used the words of laws to create connections. Legislation is normative creation – it creates mostly compulsory norms for people to live by. They theorized that a norm is a sentence in the natural language. They then also theorized that the largest semantic value of the sentence is in the verbs and nouns. They collected the word pairs in a sentence into a weighted network, where each node is a word, a noun or a verb, and each edge connects two words. Each edge has a weight equal to the number of times the two words appeared together. This novel approach allowed them to study the topological structure of the language used in law. For evaluation purposes they selected random acts from Estonian law and compared them to each other using this language network. They also built a tree of closest-connected Estonian legal acts. The language network is useful as it reveals a hidden structure within law that computers have not previously been able to easily access and study in such a large scale.

This work is different, as this is an empirical study. I focus on UK and Estonian law and their reference networks. I also make a thorough study of the topological structure of the legal acts and experimentally compare the topological structure of UK law with the impact of UK law.

3 Materials and methods

Legal acts data from the governments of the UK and Estonia was used to create the networks. I use the inner structure of a legal act to create a network. We look at the network of references from one section to another within a legal act. A node is a section and an edge is a reference from one section to another. We will look at the largest legal acts, that is, the laws that have the most edges.

A network of law can be defined three ways, depending on the direction of edges and simplification.

The first definition is the simplest: a directed graph, where each reference is transformed into a directed link from the referencing section to the referenced section.

The second definition is the same as the first, but duplicate edges are merged and loops are removed. Duplicate edges are two or more edges that have the same start node and the same end node. Loops are edges that begin and end at the same node.

The third definition is the undirected version of the second definition. In addition, we collapse together directed opposite edges (that is, edges that connect the same two nodes, but the direction of these edges was opposite in the directed graph) as they become duplicates after making them undirected.

This work will mostly use the second definition for topological structure analysis. If the direction of edges is ignored, the third definition is used.

For UK law, it was possible to scan through the catalogue of law that the British government has made available online. From those legal acts I was able to select the largest laws for analysis. This was not possible with Estonian law. The legal acts had to be manually entered into a database to select the largest laws for analysis from that sample.

For UK law, I calculated the complexity index developed by Kütt [20]. This measure is described in subsection 3.3.

Also for UK law, I compared the topological structure with the results of the Doing Business report. Doing Business is a report by the World Bank Group that tries to characterize the impact of law on making business. It is described in subsection 3.4.

The acquisition of UK law was done using a C program written by the author and Estonian law was acquired in cooperation between this program and Liiv et al's PHP program.

Title	Sections	Links
Corporation Tax Act 2010	1930	5200
Income Tax Act 2007	1460	3970
Proceeds of Crime Act 2002	719	3650
Corporation Tax Act 2009	1400	3450
Companies Act 2006	1500	2600
Income Tax (Trading and Other Income) Act 2005	937	2360
Town and Country Planning (Scotland) Act 1997	341	1990
Criminal Procedure (Scotland) Act 1995	746	1840
Financial Services and Markets Act 2000	897	1710
Town and Country Planning Act 1990	455	1690

Table 1: Largest legal acts

3.1 UK law

The British government has made their catalogue of law available on the web site `http://www.legislation.gov.uk`. It is possible to browse or search for most English laws on that web site, using a web browser. The site also has an API to search and browse laws in formats more easily understandable for a computer.

The data was acquired in 2018. It contains all the law available on the website at the time, that is 120 000 legal acts. I attempted to extract the sections of each legal act. I succeed with 3300 acts and extract 130 000 sections in total.

3% of the legal acts are successfully parsed, but those acts are most likely either so small, that they do not require their content to be distributed into sections, or not in the format of the UK Public General Acts. Therefore those laws are likely not large or not interesting for the general public.

On average each act has 1 section and on average each act with at least one section has 39 sections.

References from one section to another within a document were extracted from each section. I extract 260 000 references in total with 2 references per section on average. 54 000 sections have at least one reference and those sections have 5 references on average.

In summary, the ten largest legal acts are in table 1.

3.1.1 Data acquisition

The URL format for *legislation.gov.uk* is simple. Each legal act has a code and two numbers associated with it. The code is a classifier for the type of document, the first number is the year the document was published, and the second number is a serial number. The UK Public General Acts have the code *ukpga*, Church Measures *ukcm*, UK Statutory Instruments *uksi* and so on. The full list can be found at `http://www.legislation.gov.uk/browse`.

To get the URL of a legal act, these three identifiers are appended as directory names. This will, by default, return the whole document. For example the Data Protection Act 1998 has the URL `http://www.legislation.gov.uk/ukpga/1998/29`.

The URL for the API is almost the same, but one also appends `/data.xml`. The API URL for the same document is `http://www.legislation.gov.uk/ukpga/1998/29/data.xml`. This query returns the contents of that law document in a custom XML format.

While the previous query will return the current contents of the document, I will also need to query the past contents of the document by a date. For this I insert an ISO 8601 formatted date after the document serial code. To query the Data Protection Act 1998 in the state it was in August 10, 2008 from the API, I create this URL: `http://www.legislation.gov.uk/ukpga/1998/29/2008-08-10/data.xml`

I take advantage of this API to query a single whole law document with my program. All successful queries are cached locally.

The next step is to try to transform the document from the XML format into a directed graph where each node is a section in that document and each edge is a reference from one section to another in that document.

The XML document is already slightly structured when we begin. The contents of each section are in a tag with the attribute *id="section-x"*, where x is the section number. The text of the section is inside that tag and in the text there may be some presentational tags for list points or subsections.

After parsing the XML document, the document is divided into sections with each section containing legal text understandable for humans. We will look for references to sections within that text.

I begin by trying to find every word *section* x or *sections* x within the text, where x is a number. I then push the location into the real parser. Parsing the section number can be complicated as there are many variations.

The section number might appear as a number like "56" or as a number and a subsection number or even multiple subsection numbers such as “56(a)” or “56(a)(b)”. After the section number might come the words *and*, *or* or *to*. All of them mean that there will be another section reference. With *and* and *or* I merely need to add that second section reference. With *to* I will also need add all the implicit references between the first and second references to my list of edges. For example there may be *sections 56 or 60* or there may be *sections 56 to 60*. The words *and*, *or* or *to* may appear in any combination and any number of times.

Finally the reference might actually be to another legislative document. For consistency, references to outside documents were ignored. In that case the reference will end with the words *of The Law of Marriage* or some other legal act. If the word *of* occurs immediately after the list of references, those references will be ignored. But sometimes the reference will also end with the words *of this Act*. References with that specific phrase are not ignored.

After having parsed section references within the text for each section, we now have a directed graph of sections referencing each other within the document.

3.2 Estonian law

Liiv et al [17] built a parser of Estonian legal acts. They used it to create the same network the author has created for UK law: each node is a section and each link is a directed reference from section to section. This work uses Liiv’s shared parser to create similar networks for a handful of Estonian laws and pick the largest one.

The largest law turned out to be Law of Obligations Act 2018. It has 384 sections and 492 references.

3.3 Complexity

Sinha [21] has developed a complexity index based on chemistry. He developed a simple equation where the complexity of a system can be expressed as

$$C = C_1 + C_2 \times C_3$$

where C_1 is the inner complexities of each component, C_2 is the complexities of interactions between each component and C_3 is the complexity of the architecture of the system. They are each defined as

$$\begin{aligned} C_1 &= \sum_{i=1}^{n} \alpha_i \\ C_2 &= \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} A_{ij} \\ C_3 &= \gamma E(A) \end{aligned}$$

where n is the number of components, α_i is the complexity of the component i, β_{ij} is the interaction complexity between components i and j, A is the DSM (Design Structure Matrix) of the system, $\gamma = \frac{1}{n}$ is the scaling factor and $E(A)$ is the graph energy. The graph energy is defined as

$$E(A) = \sum_{i=1}^{n} \sigma_i$$

where σ_i is the i-th eigenvalue of the matrix. In total

$$\begin{aligned} C &= C_1 + C_2 \times C_3 \\ &= \sum_{i=1}^{n} \alpha_i + \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} A_{ij} \right) \left(\frac{E(A)}{n} \right) \end{aligned} \tag{1}$$

α_i and β_{ij} are domain-specific and are left for the implementer to define. He also does not define how to construct the network of the system and therefore, the Design Structure Matrix A.

Andres Kütt [20] has taken Sinha's work and developed definitions in the domain of law for the variables not specified by Sinha: α_i, β_{ij} and the DSM A.

> "In contrast to both abstract ideas and domain-specific solutions, Sinha provide an approach to complexity rooted in systems engineering, a field focused on general socio-technical systems. Borrowing from the field of chemistry, a complexity index is derived that has been validated by further research." [20]

Kütt constructed a three-layered approach, where each layer uses Sinha's formula with different components. At the highest layer, Kütt seeks the complexity of all of legislation in Estonia. He finds the connections between legal acts by laws referencing each other. This will form his network and his DSM A. He uses text parsing for this. Kütt defines β_{ij} as

$$\beta_{ij} = \max\left(\alpha_i, \alpha_j\right) c_{ij} \tag{2}$$

where α_i and α_j are the complexities of legal acts and c_{ij} is the number of references between the legal acts i and j. The challenge is then to obtain the individual complexities of legal acts. He uses (1) again with the same β_{ij} definition (2), but this time the components are sections and connections are references between sections. To find the complexity of the individual sections, if a section has subsections, he again uses (1) with (2) and with connections as references between subsections. If a section does not have subsections, he calculates the individual complexity using the same algorithm as for subsections. To calculate the individual complexity of subsections, he uses the subsections' legal text.

For text complexity, Kütt uses a similar approach to Kolmogorov complexity [22]: he compares the existing text with the minimal encoding of the text. If there is a lot of additional data in the text, it is unnecessarily complex. He acquires the minimal encoding of the text by using LZMA compression. To achieve more precise complexity calculation, the text should be normalized to remove variation in different word forms, like "is", "was" or "were", and different word lengths, like "car", "bicycle" or "compression". First the text is lemmatized and then each word is replaced with a random unicode character. The text complexity is defined as

$$C_m\left(t\right) = \frac{\left|L\left(\varPhi\left(t\right)\right)\right|}{\left|\varPhi\left(t\right)\right|}$$

where t is the text, $\varPhi\left(t\right)$ is the normalized text and the $L\left(x\right)$ is the LZMA compression.

3.4 Doing Business data

World Bank Group has been publishing an annual report "Doing Business" [23]. The report takes measurements of legislation in a large amount of countries to measure the impact of legislation on business. Doing Business has a lot of metrics that have been sorted into categories. The Doing Business categories are: Starting a Business, Dealing with Construction Permits, Getting Electricity, Registering Property, Getting Credit, Protecting Minority

Investors, Paying Taxes, Trading Across Borders, Enforcing Contracts, Resolving Insolvency.

Each metric also has a Distance to Frontier score.

> “The distance to frontier score shows the distance of an economy to the "frontier," which is derived from the most efficient practice or highest score achieved on each indicator.” [24]

The highest score is 100 and the lowest score is 0. The Distance To Frontier score of a category is the average Distance To Frontier score of its metrics.

World Bank Group has been publishing this information since 2004 [25] and it is available online.

World Bank Group also publishes online which laws in each country are used for their measurements [26]. The categories represented there are not all clearly mapped to Doing Business categories. Included in these results are six categories that clearly map to earlier Doing Business categories: Starting a Business, Registering Property, Getting Credit, Paying Taxes, Enforcing Contracts, Trading across Borders. It was not possible to contact World Bank Group for the missing information.

In these six categories it was possible to get a list of all the legal documents pertaining to that category, to compare them to topological structure results.

Each legal document was transformed into a graph and topological structure measures were calculated, as well as Kütt’s complexity index. The complexity of the graph should also correlate with the difficulty a human will have in understanding the document.

Five topological structure measures were used for legislation:

1. Average vertex degree
2. Average path length
3. Graph diameter
4. Graph clustering coefficient
5. Andres Kütt complexity [20]

The measures were saved in a database table for later validation.

Results of each measure for the Doing Business category Starting a Business can be seen in table 2. It shows the average measure across all the law

	Average Path Length	Average Vertex Degree	Diameter	Global Clustering Coefficient	Kütt Complexity
2004	1.57	1.24	4.00	0.09	335.69
2005	1.56	1.24	3.85	0.09	319.47
2006	1.56	1.24	3.85	0.09	319.23
2007	1.77	1.26	4.79	0.10	540.67
2008	1.75	1.25	4.71	0.10	538.17
2009	1.75	1.23	4.79	0.11	533.39
2010	1.74	1.24	4.79	0.10	540.81
2011	1.74	1.23	4.79	0.10	540.11
2012	1.74	1.24	4.79	0.10	543.46
2013	1.74	1.24	4.79	0.10	542.73
2014	1.76	1.25	4.79	0.10	560.38
2015	1.76	1.24	4.79	0.10	559.19
2016	1.76	1.23	4.64	0.10	569.39
2017	1.93	1.23	4.86	0.10	582.28
2018	1.94	1.23	4.86	0.10	585.85

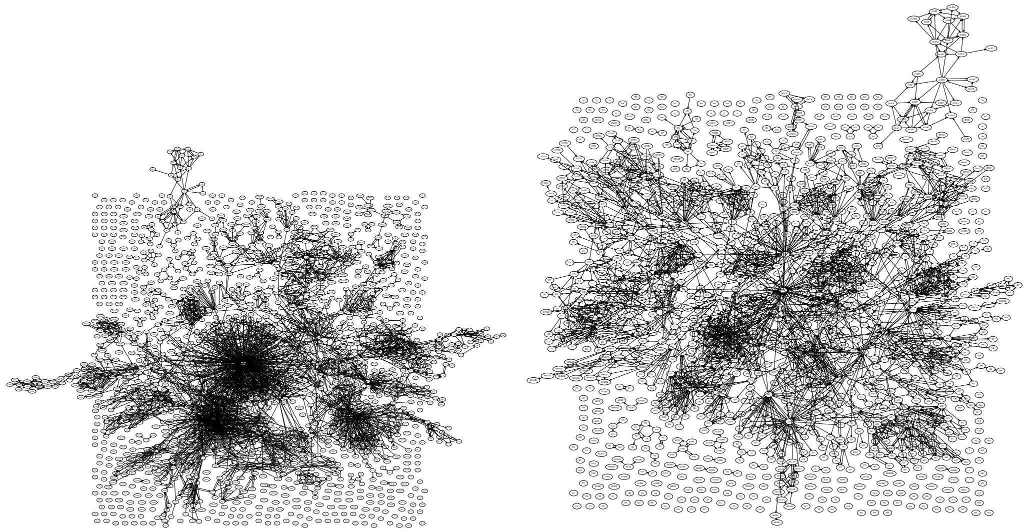
Table 2: Topological structure measures for Starting a Business category

	Average Path Length	Average Vertex Degree	Diameter	Global Clustering Coefficient	Kütt Complexity
Average Path Length	1.00	0.62	0.99	−0.32	0.89
Average Vertex Degree	0.62	1.00	0.67	0.36	0.58
Diameter	0.99	0.67	1.00	−0.31	0.89
Global Clustering Coefficient	−0.32	0.36	−0.31	1.00	−0.29
Kütt Complexity	0.89	0.58	0.89	−0.29	1.00

Table 3: Correlation of algorithms to each other

documents as they were in effect in that year as calculated by different graph measures.

I also compare all the algorithms to each other by correlation. The results can be seen in table 3.



(a) Corporation Tax Act 2010

(b) Income Tax Act 2007

Figure 1: UK legislation graphs

4 Results

The topological structure analysis and the Doing Business correlation analysis were done in R [27] and Kütt's complexity index calculation was done using Python.

For UK law, we will focus on two legal acts that had the most references: Corporation Tax Act 2010 and Income Tax Act 2007. Corporation Tax Act 2010 had 1930 sections and 5200 references while Income Tax Act 2007 had 1460 sections, 3970 references. Their directed graphs are visible in figure 1.

The large scale of the networks makes manual analysis impractical, so we use statistical methods for further analysis.

4.1 Topology structure

A path from one vertex to another is the set of edges that can be used to move from the first vertex to the last (possibly moving through other vertices). In a directed graph one is allowed to move only in the direction of the edge, in an undirected graph one may move in both directions along the edge.

The vertices of a graph can be divided into components. A component is a set of vertices that have a path from all of them to all the other vertices, i.e. two vertices are in the same component if there exists both a path from the first vertex to the second and from second to the first. In a directed graph, if edge direction matters, the component is usually called strongly connected, otherwise it is called weakly connected.

Some components of a graph are more interesting than others. The largest component of the graph is called a giant component. In a directed graph, it is named the Giant Strongly Connected Component (GSCC). The vertices that have a path to the GSCC form the Giant In-Component and vertices that have a path from the GSCC form the Giant Out-Component. Tendrils are components that have a path from the Giant In-Component or have a path to the Giant Out-Component. Detached Components (DC) are components that are completely separate.

The Giant Weakly Connected Component or the Giant Connected Component (GCC) is the largest weakly connected component. It is also the largest component in any graph. The Giant In-Component, the Giant Out-Component and the tendrils are part of the GCC.

The components of the networks of the two laws are calculated. The GCC of the Corporation Tax Act 2010 consists of 1420 nodes. 74% of all nodes are part of the GCC. The other 26% form 384 DC. The GSCC consists of 94 nodes and forms 5% of the whole network.

The Income Tax Act 2007 has a GCC of 1130 nodes meaning 77% of the network is part of the GCC. 247 DC are formed by the other 23%. The GSCC consists of 283 nodes and makes up 19% of the network.

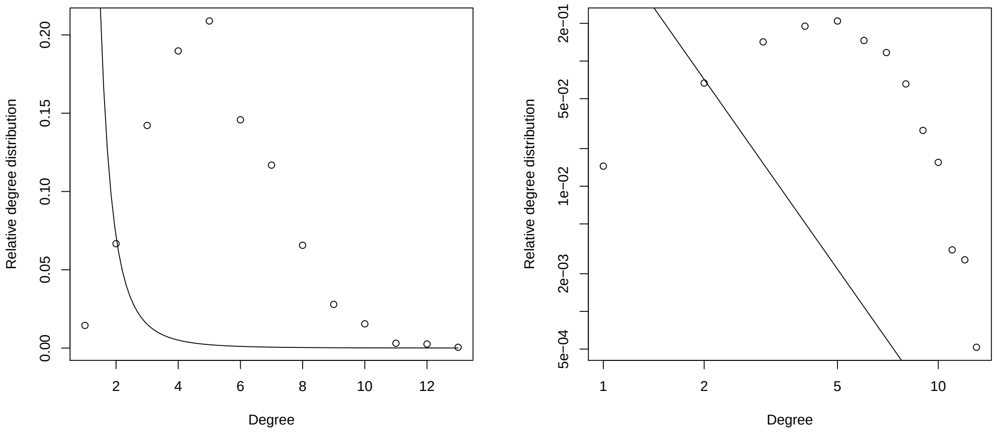
The in-degree of a vertex is the count of edges that begin from the vertex. The out-degree of a vertex is the count of edges that end at the vertex. The degree of a vertex is the sum of the in-degree and out-degree of the vertex.

The average degree of a graph is the average of the in-degree or the out-degree of the vertices:

$$\langle k \rangle = \frac{1}{n} \sum k^d = \frac{1}{n} \sum k^o = \frac{m}{n}$$

where n is the vertex count, m is the edge count, k^d is the in-degree of a vertex and k^o is the out-degree of a vertex.

For random networks the links usually follow the Poisson distribution. I created a random network with the same number of nodes and links as the Corporation Tax Act 2010 and calculated its degree distribution. Results can be seen in figure 2.



(a) Linear scaling

(b) Logarithmic scaling

Figure 2: Random network degree distribution. The points are the distribution of the network and the line is power-law function fitted to the distribution.

While random networks tend to follow a Poisson distribution, complex system network links tend to follow power-law distribution.

Power-law distribution is mathematically defined as

$$P(k) = k^{-\alpha}$$

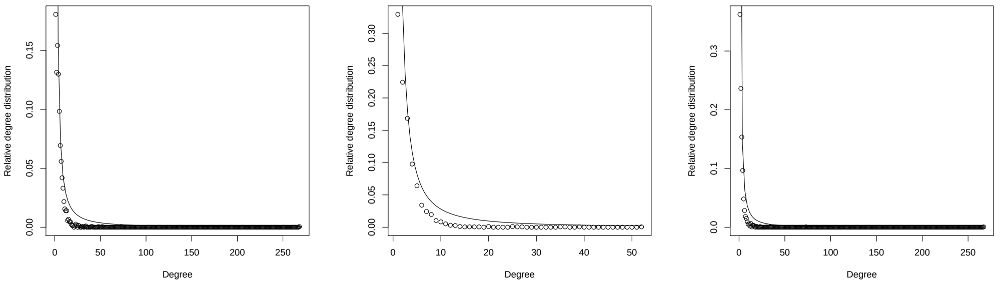
where $P(k)$ is the power-law function, k is the degree of the vector and α is the scaling exponent. System networks that follow a power-law function are called scale-free networks.

Adamic et al [28] showed that the World Wide Web follows a power-law distribution and is a scale-free network. Torre et al [8] showed that the economy of a nation also follows a power-law distribution and therefore is also a scale-free network.

The average degree of Corporation Tax Act 2010 is:

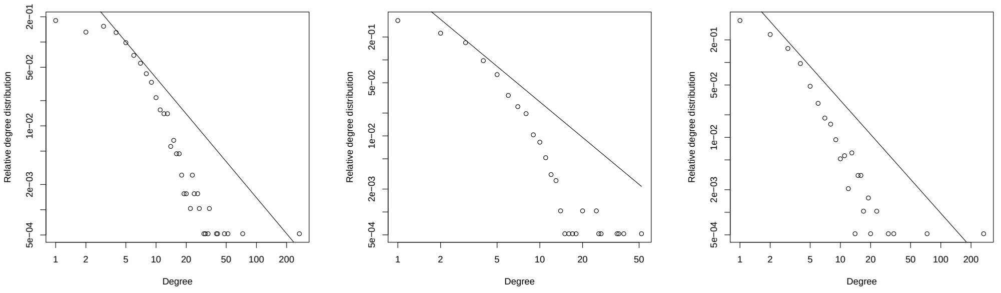
$$\langle k \rangle = 2.04$$

18% of nodes have no links and 11% of nodes have only one link. The highest degree in the network is 267. The degree distributions of Corporation Tax



(a) Degree distribution (b) In-degree distribution (c) Out-degree distribution

Figure 3: Corporation Tax Act 2010 degree distributions



(a) Degree distribution (b) In-degree distribution (c) Out-degree distribution

Figure 4: Corporation Tax Act 2010 degree distributions in logarithmic scale

Act 2010 can be seen in figures 3 and 4. On these figures there is also the power-law function fitted to the degree distribution. The power-law is fitted by finding the scaling exponent using maximum likelihood estimation.

The total degree distribution follows a power-law (figures 3a and 4a):

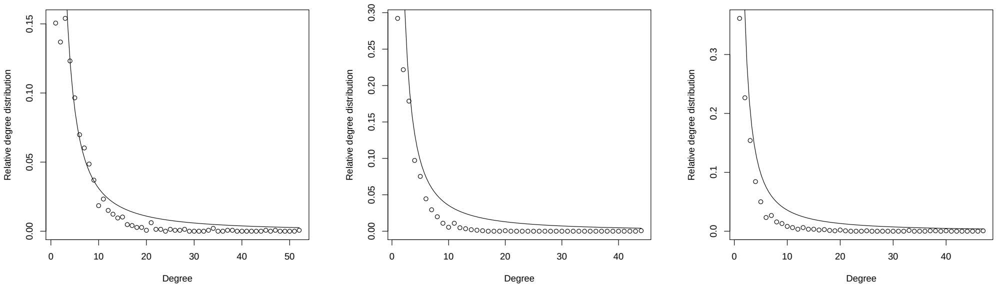
$$P(k) = k^{-1.43}$$

the in-degree distribution follows a power-law (figures 3b and 4b):

$$P(k) = k^{-1.55}$$

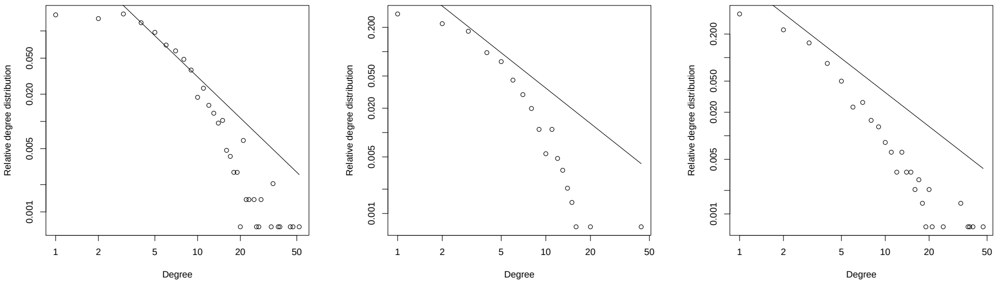
and the out-degree distribution follows a power-law (figures 3c and 4c):

$$P(k) = k^{-1.51}$$



(a) Degree distribution (b) In-degree distribution (c) Out-degree distribution

Figure 5: Income Tax Act 2007 degree distributions



(a) Degree distribution (b) In-degree distribution (c) Out-degree distribution

Figure 6: Income Tax Act 2007 degree distributions in logarithmic scale

The Income Tax Act 2007 has an average degree of

$$\langle k \rangle = 2.15$$

About 15% of the nodes in the network have no links and about 12% of the nodes have one link. The highest node degree in the graph is 51. The degree distribution can be seen in figures 5 and 6. Again, I fitted a power–law to the distribution using maximum-likelihood estimation.

From figures 5a and 6a, the total degree distribution follows a power-law

$$P(k) = k^{-1.51}$$

the in-degree distribution follows a power-law (figures 5b and 6b)

$$P(k) = k^{-1.45}$$

and the out-degree distribution fits a power-law (figures 5c and 6c)

$$P(k) = k^{-1.45}$$

Since the degree distributions of the networks of both laws follow the power-law distribution, one can deduce that they are scale-free networks.

The clustering coefficient of a graph is a measure that will show how much of a graph is formed into tightly connected clusters. It is also called transitivity or the global clustering coefficient:

$$C = \frac{\text{number of closed triples}}{\text{number of all triples}}$$

A triplet consists of three connected vertices. A triplet is open if the connection is made by two edges and closed if the connection is made by three edges forming a closed loop. Each triplet is counted for each vertex in total of three times. Both the direction of the edges and the isolated vertices with no neighbors or only one neighbor are ignored.

In essence the clustering coefficient shows the extent to which the graph forms a small world where everyone knows everyone else. A graph with a small clustering coefficient has little redundancy and is more similar to a tree.

The clustering coefficient of a vertex is a bit different. It represents how well the neighbors of a vertex are connected to each other. It is also called the local clustering coefficient:

$$\begin{aligned} C_i &= \frac{\text{existing edges between neighbours}}{\text{possible edges between neighbours}} \\ &= \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i\,(k_i - 1)} \end{aligned}$$

where N_i is the neighbor vertices of vertex v_i, E the set of all edges and k_i the degree of a vertex. Again, both the direction of the edges and isolated vertices are ignored.

The average local clustering coefficient is calculated by finding the average of the local clustering coefficients for all the non-isolated vertices:

$$\langle C\rangle = \frac{1}{n}\sum C_i$$

The average local clustering coefficient is similar, but not the same as the global clustering coefficient. The global clustering coefficient tends to add more weight towards high degree vertices and the local clustering coefficient adds more weight towards smaller degree vertices.

The global clustering coefficient of Corporation Tax Act 2010 is:

$$C = 0.0988$$

I also calculated the local clustering coefficient for each vertex. Results can be seen in figure 7a. The average of the local clustering coefficients is:

$$\langle C\rangle = 0.409$$

The global clustering coefficient of Income Tax Act 2007 is:

$$C = 0.203$$

and the local clustering coefficients can be seen in figure 7b. The average local clustering coefficient is:

$$\langle C\rangle = 0.377$$

The global clustering coefficient of a random graph is

$$C = 0.00227$$

and the average local clustering coefficient is:

$$\langle C\rangle = 0.00185$$

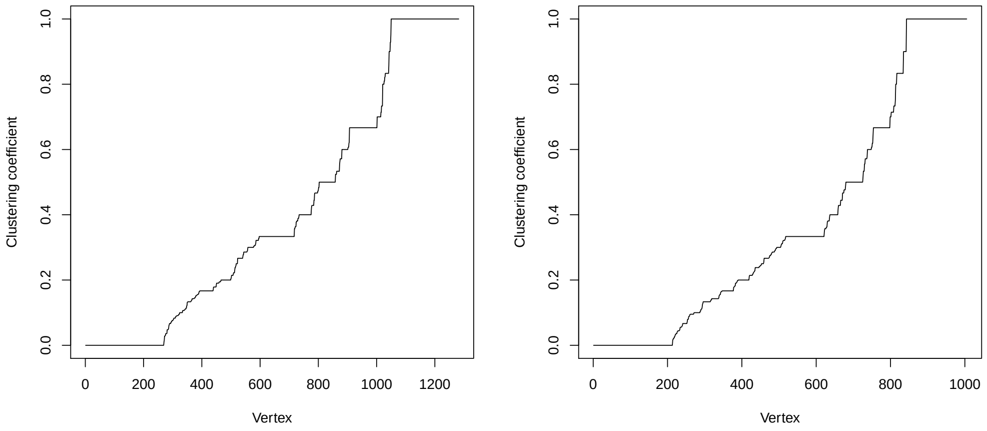
The global clustering coefficient is about 40 times lower and the average local clustering coefficient is about 180 times lower than for the network of Corporation Tax Act 2010. The law networks are not at all random.

The betweenness of a vertex is a centrality measure of how much traffic passes through the vertex. It is calculated by calculating the shortest paths that cross the vertex compared to all the shortest paths. The formula is:

$$\sigma(m) = \sum_{i\neq j}\frac{B\left(i,m,j\right)}{B\left(i,j\right)}$$

Type	Network	Exponent	Clustering coefficient
Economical	Bank of Japan payments [5]	$\gamma = 2.1$	—
	US Federal Reserve Bank [6]	$\gamma_{out} = 2.15$ $\gamma_{in} = 2.11$	0.53
	Austrian Interbank Market payments [7]	$\gamma_{out} = 3.1$ $\gamma_{in} = 1.7$	0.12
	Estonia Swedbank payments [8]	$\gamma_{out} = 2.39$ $\gamma_{in} = 2.49$	0.183
Technological	WWW [9]	$\gamma_{out} = 2.4$ $\gamma_{in} = 2.1$	—
	Peer-to-peer network [3]	$\gamma = 2.1$	0.012
	Digital electronic circuits [10]	$\gamma = 3$	0.03
Social	Film actors [11]	$\gamma = 2.3$	0.78
	Email messages [12]	$\gamma_{out} = 2.0$ $\gamma_{in} = 1.5$	0.16
	Telephone calls [13]	$\gamma = 2.1$	—
Biological	Protein interactions (yeast) [14]	$\gamma = 2.4$	0.022
	Metabolism reactions [15]	$\gamma_{out} = 2.2$ $\gamma_{in} = 2.2$	0.32
	Energy landscape for a 14-atom cluster [16]	$\gamma = 2.78$	0.073
Law (from this work)	Corporation Tax Act 2010	$\gamma_{out} = 1.51$ $\gamma_{in} = 1.55$	0.409
	Income Tax Act 2007	$\gamma_{out} = 1.45$ $\gamma_{in} = 1.45$	0.377
	Law of Obligations Act 2018	$\gamma_{out} = 1.37$ $\gamma_{in} = 1.5$	0.188

Table 4: Comparison to other large scale networks.



(a) Corporation Tax Act 2010

(b) Income Tax Act 2007

Figure 7: Growth of clustering coefficient. There are fewer nodes since isolated nodes were ignored.

where $B\left(i, m, j\right)$ is the number of shortest paths from vertex i to vertex j that crosses vertex m, $B\left(i, j\right)$ is the total number of shortest paths from i to j and the sum is over all pairs of vertices that have a path from i to j.

It is assumed that vertices with high betweenness have greater control over the graph, making them either the strongest part of the graph or the weakest part. For example, in a network of payments, high centrality means that a company is vital to many other companies, meaning it probably brings in a large profit. That company would also be the weakest part of the network, because the shutdown of this company could paralyze a large swath of other companies.

Corporation Tax Act 2010 has an average betweenness score of 300 and the Income Tax Act 2007 has an average betweenness score of 1960. This shows that the Income Tax Act 2007 is more structured, less chaotic and easier to read.

Edge betweenness is identical to vertex betweenness except that instead of measuring how much a vertex is in the shortest paths, it measures how much an edge is in the shortest paths. The formula is identical to the one above, except m is now an edge instead of a vertex.

The edge betweenness of Corporation Tax Act 2010 is 172 and the edge betweenness of Income Tax Act 2007 is 1020.

The average shortest path of a graph is the average of shortest paths between all the pairs of vertices in the graph:

$$\langle l \rangle = \frac{1}{n} \sum_{m \neq n} l_{m,n}$$

where $l_{m,n}$ is the shortest path between vertices m and n.

The average shortest path of Corporation Tax Act 2010 is 7.03 and the average shortest path of Income Tax Act 2007 is 9.73. On average every node is about 8 steps away from any other node.

The diameter of a graph is the longest shortest path in the graph:

$$d = \max_{m,n} l_{m,n}$$

Corporation Tax Act 2010 has a diameter of 21 and the diameter of Income Tax Act 2007 is 27.

The eccentricity of a node is the longest shortest path to any other node and the radius of a graph is the smallest eccentricity:

$$r = \min_{m} \max_{n} l_{m,n}$$

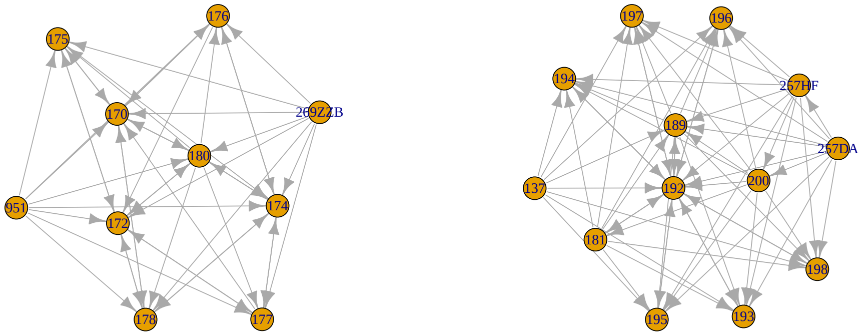
One could have also defined diameter as having the largest eccentricity of the graph:

$$d = \max_{m} \max_{n} l_{m,n}$$

Since the eccentricity of isolated vertices will be 0, the radius of a graph with isolated vertices will be 0. Since that is not relevant, I measure the radius of the GCC and the GSCC.

For a directed graph, the eccentricity of a vertex can be calculated in three ways: ignoring the direction of edges for the shortest paths, calculating only the shortest paths that are inbound for the vertex and calculating the shortest paths that are outbound from the vertex.

As there are three types of eccentricities, there are three types of radiuses: undirected radius r, out-radius r_{out} and in-radius r_{in}. r_{in} and r_{out} are likely to be the same.



(a) Corporation Tax Act 2010

(b) Income Tax Act 2007

Figure 8: k-core

The undirected radius of the GCC Corporation Tax Act 2010 is $r = 10$, the out-radius is $r_{out} = 8$ and the in-radius is $r_{in} = 8$. Income Tax Act 2007 has a undirected radius of $r = 8$, an out-radius of $r_{out} = 11$ and an in-radius of $r_{in} = 10$. The larger undirected radius comes from the GCC being larger than GSCC.

The k-core of a graph is the maximal sub graph where each vertex has k degree. The vertex with the smallest degree is iteratively removed until only vertices with k degree or more remain.

The k-core of a vertex is defined as belonging in the k-core of a graph, but not in the (k+1)-core of the graph.

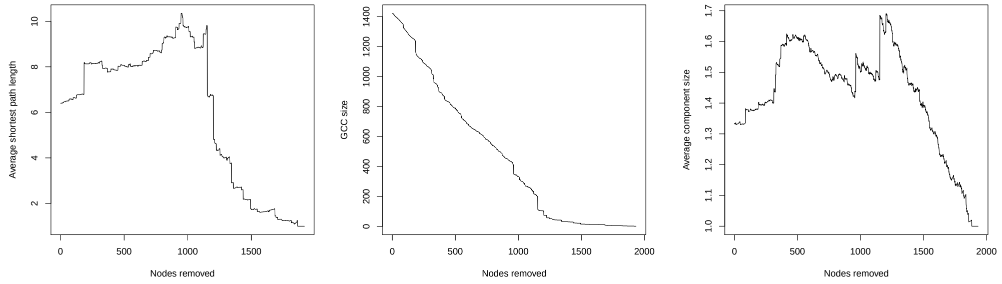
Corporation Tax Act 2010 has 9 k-core levels, from 0 to 8. The last 8-core has 10 elements. The resulting network can be seen in figure 8a. Income Tax Act 2007 has also 9 k-core levels, from 0 to 8. The last level has 13 elements. The 8-core can be seen in figure 8b.

4.1.1 Robustness simulation

Another interesting aspect of a graph is what would happen if one would start removing nodes or edges from the graph. It is meant to represent a malicious attack, accidents or mistakes. It is also interesting to see how

Statistic	Corporation Tax Act 2010	Income Tax Act 2007
Nodes	1930	1460
References	5200	3970
Undirected links	3940	3140
$\langle k \rangle$	2.04	2.15
γ	1.43	1.51
γ_{out}	1.51	1.45
γ_{in}	1.55	1.45
C	0.0988	0.203
$\langle C \rangle$	0.409	0.377
$\langle l \rangle$	7.03	9.73
d	21	27
r	10	8
r_{out}	8	11
r_{in}	8	10
$\langle \sigma \rangle$(nodes)	300	1960
$\langle \sigma \rangle$(links)	172	1020
GCC	1420	1130
DC	384	247
GSCC	94	283
Cutpoints	222	170
Biconnected components	393	286
k-core elements	10	13
Complexity [20]	4860	3770

Table 5: Summary of statistics, where $\langle k \rangle$ is the average degree, γ, γ_{out}, γ_{in} are the undirected, out-directed and in-directed scale of the power-law function of the degree distribution, C is the global clustering coefficient or global transitivity, $\langle C \rangle$ is the average local clustering coefficient or average local transitivity, $\langle l \rangle$ is the directed average shortest path length, d is the graph diameter, r is the radius of GCC, r_{out} is the out-radius of the GSCC, r_{in} is the in-radius of the GSCC, $\langle \sigma \rangle$ is the average betweenness for the nodes and links respectively, GCC is the Giant Connected Component or Giant Weakly Connected Component, DC is Detached Components, GSCC is the Giant Strongly Connected Component.



(a) Average shortest path length

(b) GCC size

(c) Average component size excluding GCC

Figure 9: Random removal from Corporation Tax Act 2010

many nodes are vital to the structure of the graph and which ones they are. This is usually called robustness.

One way to measure robustness is to look at the components of a graph. By removing nodes, a component might split into two or more components. The removed vertices are called articulation points, cut vertices or cutpoints and they are interesting as vulnerable or key points. In this case we are observing the graph as undirected.

The Corporation Tax Act 2010 has 222 articulation points and the Income Tax Act 2007 has 170 articulation points.

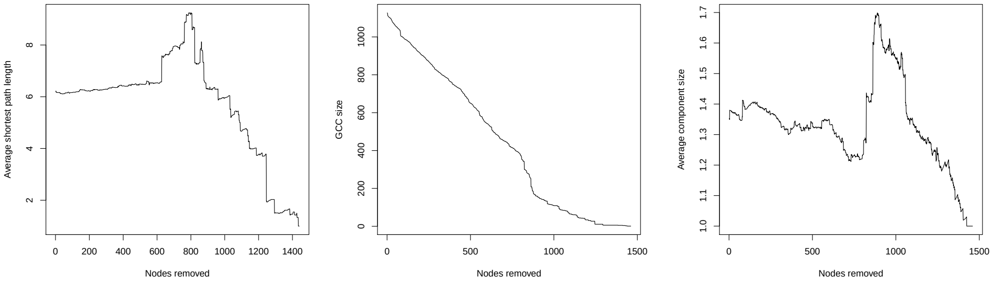
Another way to measure the robustness of a graph is to simulate the removal of nodes. One can select the nodes to remove randomly or by targeting the highest value nodes. In this measure we are dealing with an undirected graph as well.

First, let us experiment with removing nodes from the networks randomly.

Figure 9 shows the effect on the network of Corporation Tax Act 2010 when I remove random nodes. Figure 10 shows the same experiment on Income Tax Act 2007.

The networks more or less follow the pattern of scale-free networks. The average shortest path remains the same until the end when it drops sharply, the average component size has the same behavior and the GCC size slowly diminishes [29].

A second experiment is made by not selecting the removed nodes randomly but instead picking nodes with the highest betweenness. This is a



(a) Average shortest path length

(b) GCC size

(c) Average component size excluding GCC

Figure 10: Random removal from Income Tax Act 2007

good simulation of a targeted attack. Results can be seen in figure 11 for Corporation Tax Act 2010 and in figure 12 for Income Tax Act 2007.

Here the graphs again follow the same pattern as scale-free networks. The average shortest path rises sharply and then drops, so does the average component size as the GCC sharply drops.

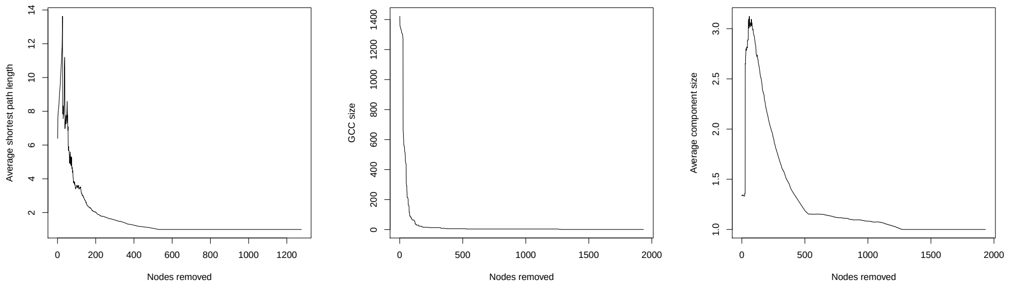
The plots do not quantify how fast the structure of the network falls apart. A simple measure is to see how many nodes must be eliminated to destroy the GCC. This is called the percolation threshold. As the GCC size drop slows down towards the end, it is interesting to note when most of the GCC has been destroyed. For that we devise the percolation threshold definition: it is the moment where the size of the GCC is at 10% of the number of vertices of the network:

$$GCC\,(p_c) = 0.1n$$

where n is the number of nodes in the graph and p_c is the percolation threshold.

For random node elimination, the Corporation Tax Act 2010 has a percolation threshold of 1150. After having eliminated 1150 nodes, the average shortest path is 6.75 and the average component size is 1.69. The Income Tax Act 2007 has percolation threshold 920. At that point the average shortest path is 6.28 and the average component size is 1.6.

When important nodes are targeted for elimination, the percolation threshold of Corporation Tax Act 2010 is 68. At percolation threshold, the average

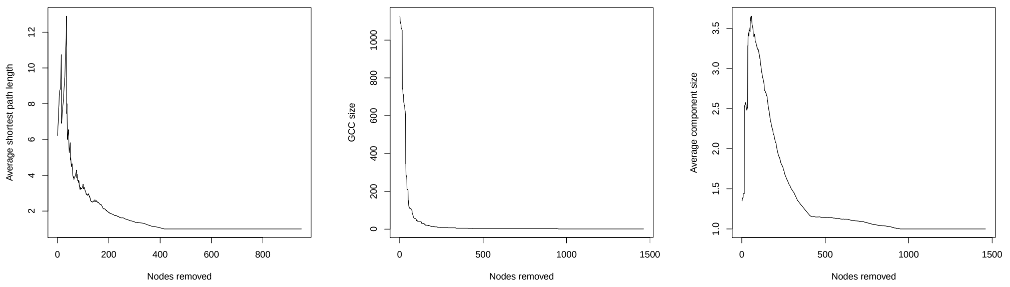


(a) Average shortest path length

(b) GCC size

(c) Average component size excluding GCC

Figure 11: Targeted removal from Corporation Tax Act 2010



(a) Average shortest path length

(b) GCC size

(c) Average component size excluding GCC

Figure 12: Targeted removal from Income Tax Act 2007

shortest path is 4.87 and the average component size is 3.06. Percolation threshold for Income Tax Act 2007 is 53 with an average shortest path of 4.72 and an average component size of 3.62.

Another important measure of a graph is the degree affinity of the nodes: are the nodes with high degree connected with other nodes with high degree and vice versa, or are nodes with high degree connected with low degree nodes and vice versa. A good way to measure this is with the average nearest neighbor degree function. In this case I am also handling the graph as undirected.

The average nearest neighbor degree function gives the average degree of a node's neighbors. It is also defined to give the average degree of the neighbors of nodes for a given degree. We are more interested in the second definition. It is defined as:

$$\langle k_{nn} \rangle (k) = \sum_{k'} k' P(k'|k)$$

where k is a degree, k' is another degree and $P(k'|k)$ is the conditional probability that a node with degree k is connected with a node with degree k' and

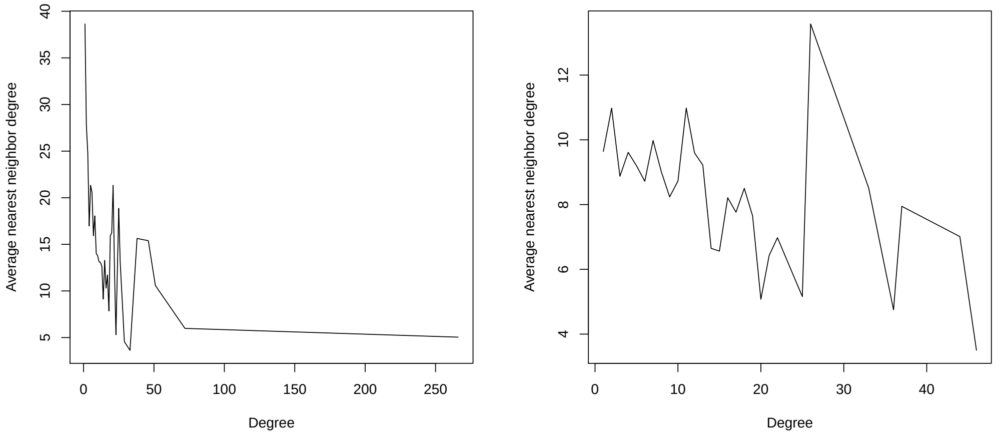
$$P(k'|k) = \frac{P(k', k)}{P(k)}$$

where $P(k', k)$ is the probability that a node with degree k' is connected to a node with degree k and $P(k)$ is the probability distribution of a node having degree k. $P(k)$ is the same as the degree distribution.

Calculating the average nearest neighbor degree function for every degree in a graph gives us an affinity distribution. This was calculated for Corporation Tax Act 2010 and Income Tax Act 2007 and the results can be seen in figure 13.

From the figures one can see that as the degree of a node gets higher, it is more likely to be connected to lower degree nodes. If a network's high degree nodes are more likely to be connected with high degree nodes, the network exhibits assortative mixing, and if a network's high degree nodes are more likely to be connected to low degree nodes, the network exhibits disassortative mixing.

Previous studies [8, 29] have shown that social networks exhibit assortative mixing and other system networks like financial, technical or financial networks exhibit disassortative mixing. To quantify this, we calculate



(a) Corporation Tax Act 2010

(b) Income Tax Act 2007

Figure 13: Average nearest neighbor degree distribution

the Pearson correlation coefficient between a degree and its average nearest neighbor degree for all degrees.

Corporation Tax Act 2010 has a degree correlation of -0.397 and Income Tax Act 2007 has a degree correlation -0.542. Networks with disassortative mixing have been shown to be fragile to targeted attacks.

In conclusion, UK legal networks are strong against random mistakes and fragile to a malicious policymaker. In terms of readability it is easier for a human to reason about such networks and as a result these legal acts are less complex when compared to social networks.

4.1.2 Complexity

A recent important topic is the complexity of law and its changes across time. As our lives get more regulated, the complexity and quantity of law only seems to increase. It would be interesting to calculate complexity for UK laws.

Let us use Kütt's algorithm [20] to calculate the complexities of UK laws. While Kütt used a three-layered scheme to calculate complexity for the entirety of legislation, we only need a single layer. We want to calculate a complexity for each individual document of law and only use the references between sections and the section's text as information. Therefore we can more simply calculate the complexity of each section's text and the system complexity of the legal act. This was done with Corporation Tax Act 2010 and Income Tax Act 2007.

Corporation Tax Act 2010 has a complexity of 4860 and Income Tax Act 2007 has a complexity of 3770.

4.1.3 Estonian law

Law of Obligations Act 2018 is used as the case study of Estonian legal acts because of its large size. Law of Obligations Act 2018 has been visualized as a graph in figure 14a. Law of Obligations Act 2018 has 384 nodes and 492 references. In the simplified graph used for analysis, the references become 390 links. This legal act was run through the same measures used for English laws.

The GCC of Law of Obligations Act 2018 consists of 149 nodes and the legal act has 76 DCs. In comparison the GSCC consists of 11 nodes.

The average degree of the graph is 1.02 and the largest degree is 26. The degree distribution seems to follow a power-law. The degree distributions are visualized in figure 15.

The scaling exponents of the power-law function for Law of Obligations Act 2018 are for the total degree

$$P(k) = k^{-1.79}$$

for the out-degree

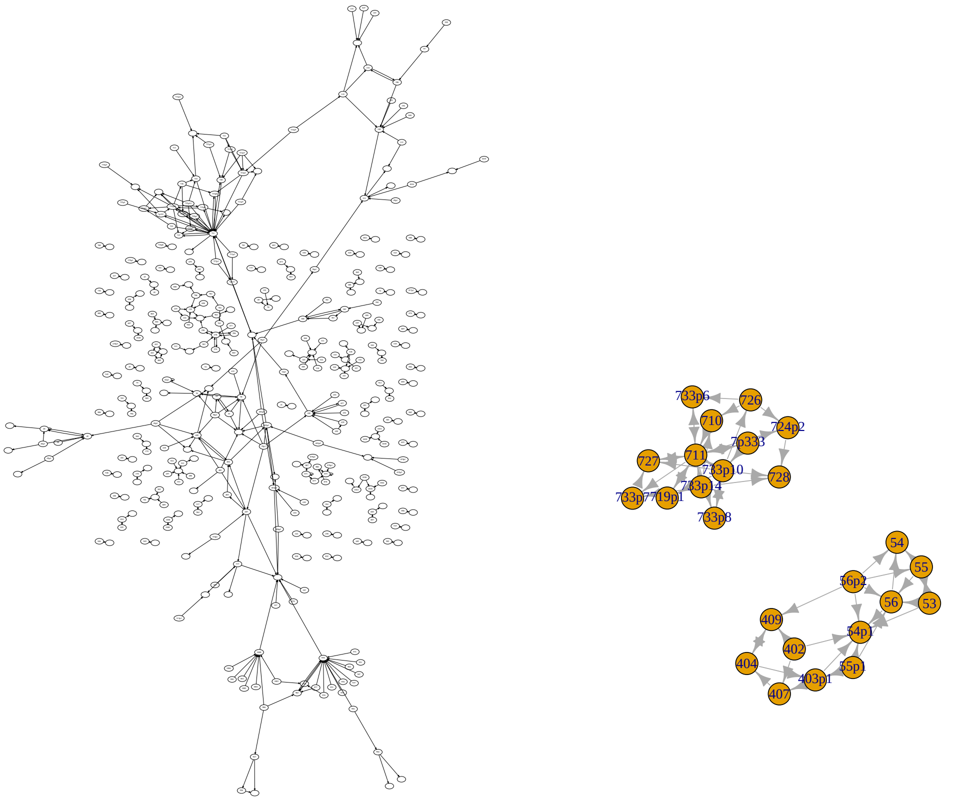
$$P(k) = k^{-1.37}$$

and for the in-degree

$$P(k) = k^{-1.5}$$

The global clustering coefficient for Law of Obligations Act 2018 is 0.12 and the local clustering coefficient is 0.188. The local clustering coefficient distribution is visualized in figure 16a.

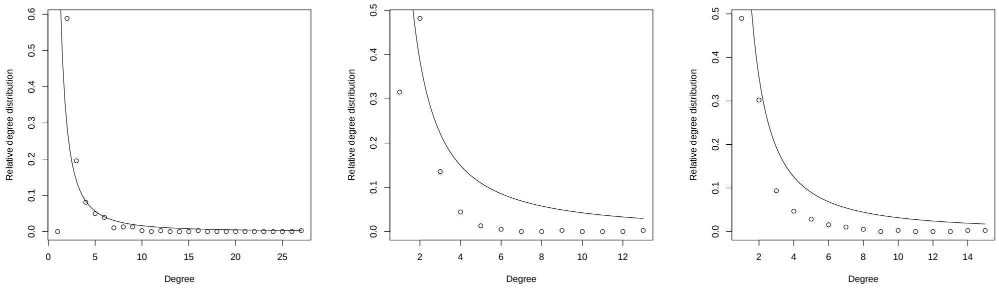
The average shortest path is 2.06 and the diameter is 6. The undirected radius for Law of Obligations Act 2018 is 7, the out-directed radius is 2 and the in-directed radius is 2.



(a) Law of Obligations Act 2018

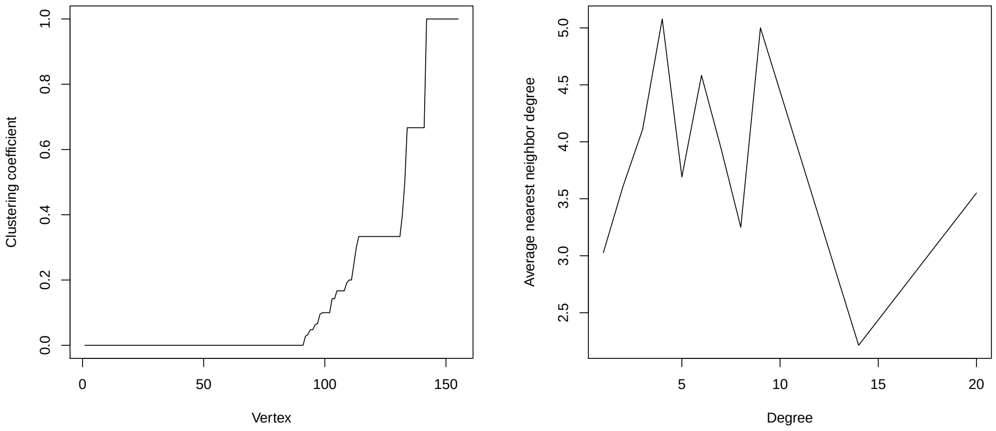
(b) Law of Obligations Act 2018 k-core

Figure 14: Law of Obligations Act 2018



(a) Total degree distribution (b) Out-degree distribution (c) In-degree distribution

Figure 15: Law of Obligations Act 2018 degree distributions



(a) Clustering coefficient

(b) Degree correlation

Figure 16: Law of Obligations Act 2018 clustering coefficient and degree correlation

The average node betweenness of the network is 3.04 and the average edge betweenness is 5.81.

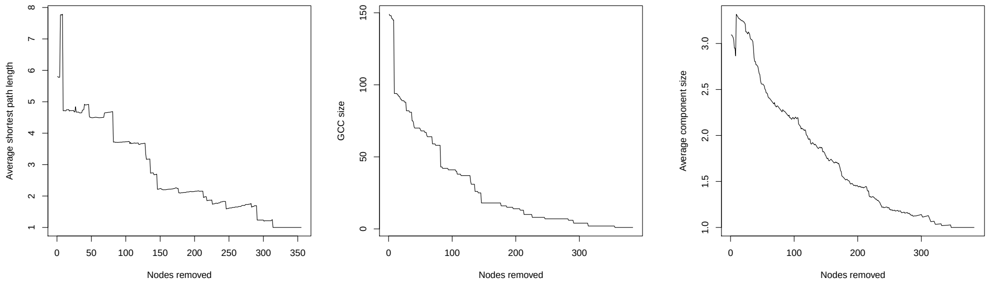
The highest k of the k-core for Law of Obligations Act 2018 is 3 and its size is 25 nodes. Interestingly enough, the k-core forms two completely separated components. The figure can be seen in 14b.

There are 95 articulation points in the graph that when removed will form a new detached component in the graph and 229 biconnected components.

When run through a robustness simulation, the percolation threshold for Law of Obligations Act 2018 is 108 nodes for random removal and 8 for targeted removal by betweenness centrality. I visualized the results of both removal strategies on figures 17 and 18.

The degree correlation is -0.262. I visualized the degree correlation in figure 16b.

It was not possible to calculate the complexity for Law of Obligations Act 2018 because of the limitations of the Liiv et al's parser.

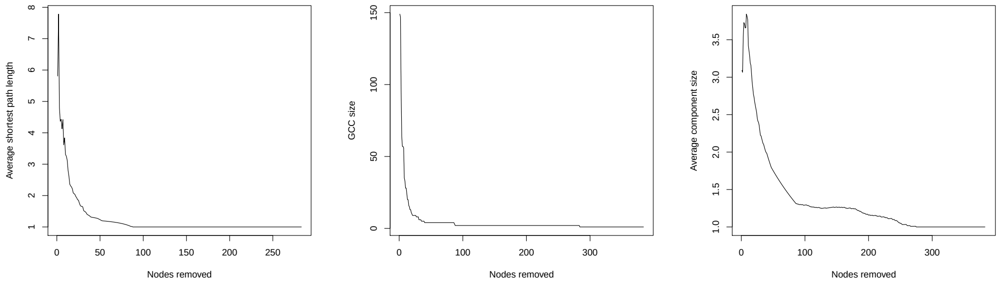


(a) Average shortest path length

(b) GCC size

(c) Average component size excluding GCC

Figure 17: Law of Obligations Act 2018 random removal



(a) Average shortest path length

(b) GCC size

(c) Average component size excluding GCC

Figure 18: Law of Obligations Act 2018 targeted removal

Algorithm 1 R data analysis of complexity to Distance to Frontier

```
GetCorrelationsToDtfByAlgorithm <-
  function(complexities) {
    correlations <-
      t(do.call(
        rbind,
        by(complexities,
           complexities$algorithm,
           function(data)
             by(data,
                data$dbu_category,
                GetCorrelationToDtf))
      ))
    rbind(
      correlations,
      "Total␣correlation" = by(complexities,
         complexities$algorithm,
         GetCorrelationToDtf)
    )
  }

GetCorrelationToDtf <- function(data) {
  cor(data$complexity, data$dtf)
}
```

4.2 Comparison to Doing Business

As an additional interdisciplinary experiment, the impact of law and the topological structure of law is compared. Using the legal acts that the World Bank Group utilized for their Doing Business report, I calculated some statistical network metrics for the networks of these laws and compared these metrics with the Distance To Frontier numbers in the Doing Business report, with the goal to explore whether Kütt's complexity [20] correlates in some form to the Distance To Frontier numbers.

A small program in R was written for analysis. The Distance To Frontier correlation analysis can be seen in algorithm 1.

	Average Path Length	Average Vertex Degree	Diameter	Global Clustering Coefficient	Kütt Complexity
Enforcing Contracts					−0.40
Getting Credit	−0.49	0.57	−0.20	−0.20	−0.56
Paying Taxes	−0.71	−0.43	−0.69	0.56	−0.63
Registering Property	−0.29	−0.44	−0.33	−0.63	−0.49
Starting a Business	0.72	−0.57	0.36	0.39	0.49
Trading across Borders	−0.77	−0.85	−0.77	−0.75	−0.88
Total correlation	0.42	0.72	0.45	0.12	0.52

Table 6: Correlation by category. The correlation of Enforcing Contracts for most algorithms is empty because the topological structure of the law doesn't change in the time period and therefore correlation calculations has no meaning.

The program queries the results coupled to previous complexity results. They were split by algorithm and Doing Business category. Finally we find the correlation between the two variables. As a final test, the data is split only by algorithm before scaling and calculating correlation. The result can be seen in table 6.

Surprisingly the total correlation is completely different from the mean across categories. The reason for this is that while the used algorithms somewhat agree with Distance To Frontier about changes in time, they don't agree with Distance To Frontier changes across categories.

These results are inconclusive – no clear pattern emerges. The Distance To Frontier represents whether the effect of the law is successful while the topological structure measures represent the complexities of law. A negative correlation was expected to show that as the effect of a law gets better, it also

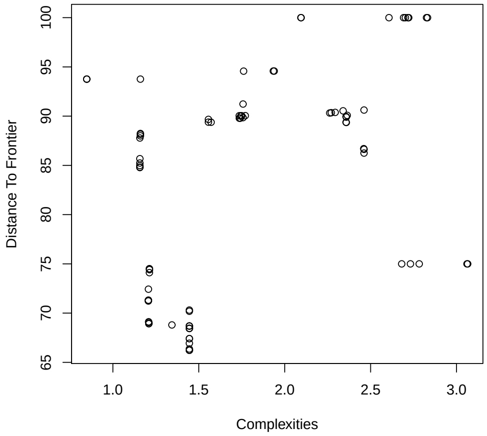


Figure 19: Distance To Frontier comparison to average vertex degree

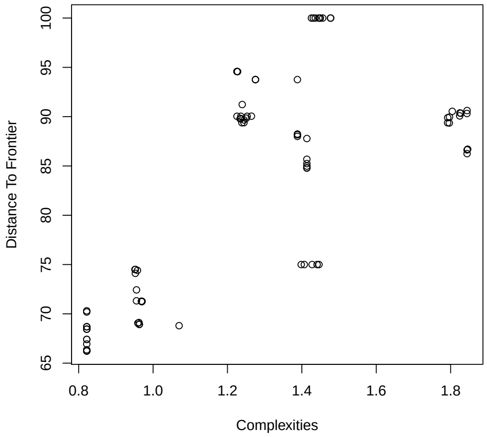


Figure 20: Distance To Frontier comparison to average path length

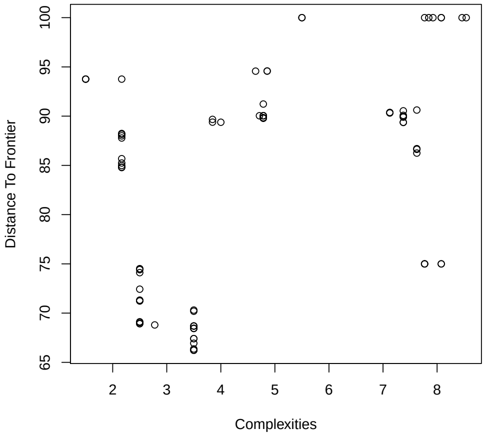


Figure 21: Distance To Frontier comparison to graph diameter

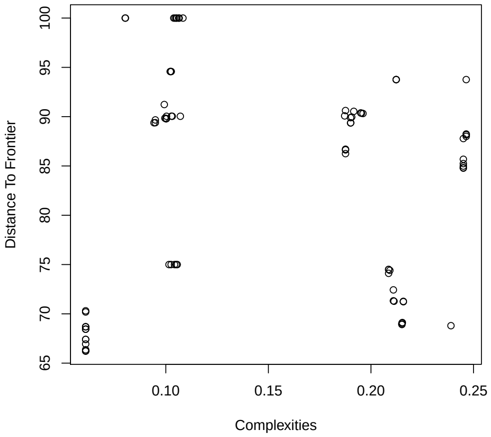


Figure 22: Distance To Frontier comparison to graph clustering coefficient

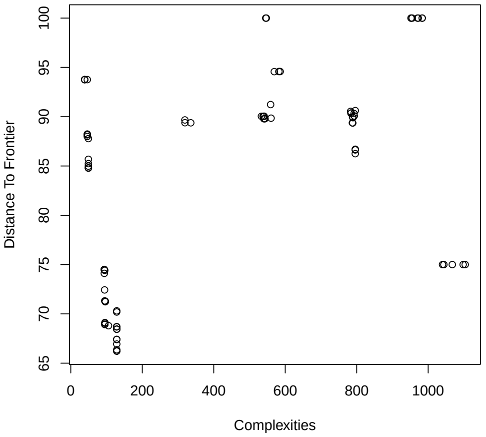


Figure 23: Distance to Frontier to Andres Kütt complexity

becomes less complex. This result can be explained by the fact that the Doing Business report only measures the impact of law and not the document of law itself and there is no correlation between the two or that as the impact of law gets better over time, the law itself seems to be only getting more complex. But even the second hypothesis is hard to confirm, as no such clear pattern emerges.

4.3 Discussion

This work compared the reference networks of two UK legal acts and one Estonian legal act. First we will discuss the differences and similarities between the UK legal acts. Then we will compare the topological structure of the domain of law to other domains. Finally we will compare Estonian and UK law.

The results are summarized in table 5. The laws were mostly similar, but surprisingly there were a few notable differences. While Corporation Tax Act 2010 was bigger and Income Tax Act 2007 was smaller, Income Tax Act 2007 has a larger GCC and GSCC. There are less DCs, cutpoints and biconnected components for Income Tax Act 2007 as well. It's betweenness is a lot larger with 1960 compared to Corporation Tax Act 2010 betweenness of 300. There seems to be a larger difference between the global clustering coefficient and the average local clustering coefficient for Corporation Tax Act 2010. As the global clustering coefficient seems to give stronger weight for nodes with high degrees and with the previously mentioned statistics, this leads to the conclusion that Corporation Tax Act 2010 has a stronger core, while Income Tax Act 2007 has its connections more spread out. In light of this it is surprising that Kütt's complexity gives a higher score for Corporation Tax Act 2010 as structures that are more focused with more independent pieces are easier for humans to read and understand.

The two UK legal acts are similar in that they both follow a power-law function with their degree distribution and are similar in other ways to scale-free networks. They have a fairly small average shortest path. When compared by removing nodes with targeted removal or random removal, targeted removal is noticeably more efficient in breaking apart the network. While the average shortest path is fairly small, it does not seem small enough for this network to be a small world network.

It makes sense that this network is scale-free. Older sections are used as references in newer sections. As previously noted, it is also easier for humans to understand text if it references either well known independent material or no material. This tendency probably brings about a scale-free, hierarchical and tree-like structure.

What is also interesting is that for both laws the global clustering coefficient was smaller than the average local clustering coefficient. We can conclude that the highly connected nodes are separated from each other. This is also logical, as it improves the readability of law.

One can see power-law scaling exponents and clustering coefficients in other domains in table 4. The domain of law differs from other domains in that law is both organically growing as well as made and organized by humans. Recently there has been more effort in reducing the complexity of law by codification and simplification. However, statistics show that law is still fairly complex. A fairly large clustering coefficient, large shortest path and a small power-law scaling exponent demonstrate that there is still a long way to go. When looking at table 4, every other domain has a scaling exponent that is clearly larger and only two other domains, film actors [11] and inter-bank payments in US [6], have a larger clustering coefficient. Those two domains are organic.

Comparing Estonian law to British law, the first thing to notice is that Estonian law is a lot smaller. This may have been caused by differences in the amount of selectable law for UK and Estonia. Note that selectable

Estonia law is small because of the deficincies in the parser. There may be other large Estonian laws that could not be found manually.

It is also clear that Law of Obligations Act 2018, like Income Tax Act 2007 and Corporation Tax Act 2010, is a scale-free network. It has a degree distribution following a power-law. It also behaves similarly under robustness simulation. Law of Obligations Act 2018 is also disassortative, like the UK law.

Interestingly, the k-core of Law of Obligations Act 2018 separates into two clusters. This may mean that there are two complicated topics in Law of Obligations Act 2018. Another measure that is different about Estonian law is the clustering coefficient. The global clustering coefficient and the average local clustering coefficient are very small compared to Income Tax Act 2007 and Corporation Tax Act 2010. That may mean that Estonian law is more structured or hierarchical.

The average node and edge betweenness is also smaller for the Estonian law document. So is the average shortest path and diameter. These numbers may result from Law of Obligations Act 2018 being smaller than its English counterparts.

5 Conclusion

This work measured the topological structure of UK and Estonian legislation, using the law that the government of the United Kingdom and Estonia published online. After obtaining a list of all legislation online and gathering the contents of every legal act, we calculated the network for each legal act, followed by statistical measures for the two UK legal acts and one Estonian legal act with the largest networks. We first analyzed the UK legal acts and then the Estonian act.

The average degree for Corporation Tax Act 2010 and Income Tax Act 2007 are 2.04 and 2.15 respectively. The degree distribution of the two networks were visualized, concluding that they follow a power-law function. We calculated the scale of the power-law function for both legal acts. The undirected power-law scaling exponent for Corporation Tax Act 2010 is 1.43 and for Income Tax Act 2007 is 1.51. We concluded that the legal acts are scale-free networks.

We calculated the average shortest path, diameter and clustering coefficient for the legal documents. The diameter for Corporation Tax Act 2010 and Income Tax Act 2007 is 21 and 27 respectively. The clustering coefficient is 0.0988 for Corporation Tax Act 2010 and 0.203 for Income Tax Act 2007. The average shortest path for Corporation Tax Act 2010 is 7.03 and for Income Tax Act 2007 is 9.73.

Kütt's [20] algorithm was used to calculate the complexity of law. Corporation Tax Act 2010 has a complexity of 4860 and Income Tax Act 2007 has a complexity of 3770.

The robustness of a network, i.e. its ability to stay intact, was measured with simulation. We simulated attacks on the network by either randomly removing nodes or selectively removing nodes with the highest betweenness centrality measure. We calculated the percolation threshold after which the GCC can be considered destroyed. The percolation threshold of a random attack for Corporation Tax Act 2010 is 1150 nodes removed and for Income Tax Act 2007 is 920 nodes removed while for a targeted attack, the percolation threshold is 68 nodes removed for Corporation Tax Act 2010 and 53 nodes removed for Income Tax Act 2007. It can be concluded that network laws are durable to random attacks or mistakes, but weak to malicious actors.

We visualized the degree correlation between neighbor nodes and calculated the correlation using Pearson correlation coefficient. For Corporation Tax Act 2010 and Income Tax Act 2007, the correlation was -0.397 and -0.542

respectively. I concluded that the networks are disassortative.

We then made the same network of references for one large Estonian law using Liiv et al's [17] parser. We found that the Estonian law is also scale-free and disassortative. The Estonian law was however much smaller and possibly more structured with a smaller clustering coefficient of 0.12.

We then tried to compare the UK networks with the results of the World Bank Doing Business report. The overall correlations between the topological structure results and the Doing Business results were 0.72 for the average degree, 0.42 for the average shortest path, 0.45 for the graph diameter, 0.12 for the the graph clustering coefficient and 0.52 for Kütt's complexity. In general no clear pattern seemed to form between the statistical network measures and the Doing Business results.

The main contributions of this thesis were:

- acquisition of the data, that is, examples of the UK and Estonian law;
- parsing of references in the UK and Estonian law;
- a thorough statistical analysis of the network of references;
- as an additional interdisciplinary experiment, a comparison of the topological structure of UK networks to the results of the Doing Business report.

5.1 Limitations and future research

The parser of UK documents and the online database `http://www.legislation.gov.uk/` had some limitations. Not all documents were current. Nor were all documents in the same format. Some documents were only pdf format, which was unreadable for my parser. Therefore some legal acts were left out. The reference parser was not completely accurate. Sometimes the language used was confusing for the parser and the author. Some false references may have been put in the network.

Liiv et al's Estonian law parser was limited in that it left out sections that had no references and didn't return the section texts. This may have skewed the statistical measurements and calculation of Kütt's complexity index [20] was impractical. This may have influenced the comparison between Estonian and UK law.

These are previously unstudied complex networks. As the parser is open-source, the experiments can be easily reproduced after refinement.

The UK and Estonian parser could be modified to recognize only references between documents. This would produce a single network of the whole UK law or the whole Estonian law that could then be analyzed.

By modifying the UK parser, one could study the change of topological structure in time.

References

[1] S. D. Eppinger and T. R. Browning, *Design Structure Matrix Methods and Applications.* MIT press, 2012.

[2] D. Bu, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu, J. Zhang, S. Sun, L. Ling, N. Zhang, G. Li, and R. Chen, “Topological structure analysis of the protein-protein interaction network in budding yeast,” *Nucleic Acids Research*, vol. 31, no. 9, p. 2443–2450, May 2003. [Online]. Available: http://dx.doi.org/10.1093/nar/gkg340

[3] M. Ripeanu, I. Foster, and A. Iamnitchi, “Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design,” *arXiv preprint cs/0209028*, 2002.

[4] G. Fagiolo, J. Reyes, and S. Schiavo, “On the topological properties of the world trade web: A weighted network analysis,” *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 15, p. 3868–3873, Jun 2008. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2008.01.050

[5] H. Inaoka, T. Ninomiya, K. Tanigushi, T. Shimizu, and H. Takayasu, “Fractal network derived from banking transaction,” *An analy'sis of network structures formed by financial institutions. Bank of Japan Work'ing Paper Series*, no. 04, 2004.

[6] K. Soramäki, M. L. Bech, J. Arnold, R. J. Glass, and W. E. Beyeler, “The topology of interbank payment flows,” *Physica A: Statistical Mechanics and its Applications*, vol. 379, no. 1, p. 317–333, Jun 2007. [Online]. Available: http://dx.doi.org/10.1016/j.physa.2006.11.093

[7] M. Boss, H. Elsinger, M. Summer, and S. Thurner 4, “Network topology of the interbank market,” *Quantitative finance*, vol. 4, no. 6, pp. 677–684, 2004.

[8] S. Rendón de la Torre, J. Kalda, R. Kitt, and J. Engelbrecht, “On the topologic structure of economic complex networks: Empirical evidence from large scale payment network of estonia,” *Chaos, Solitons & Fractals*, vol. 90, p. 18–27, Sep 2016. [Online]. Available: http://dx.doi.org/10.1016/j.chaos.2016.01.018

[9] R. Albert, H. Jeong, and A.-L. Barabási, "Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, p. 130–131, Sep 1999. [Online]. Available: http://dx.doi.org/10.1038/43601

[10] R. Ferret, R. V. Cancho, and R. V. Solé, "Statistical mechanics of complex networks," *Lecture Notes in Physics*, 2003. [Online]. Available: http://dx.doi.org/10.1007/b12331

[11] D. J. Watts and S. H. Strogatz, "Collective dynamics of "small-world" networks," *Nature*, vol. 393, no. 6684, p. 440–442, Jun 1998. [Online]. Available: http://dx.doi.org/10.1038/30918

[12] H. Ebel, L.-I. Mielsch, and S. Bornholdt, "Scale-free topology of e-mail networks," *Physical Review E*, vol. 66, no. 3, Sep 2002. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.66.035103

[13] W. Aiello, F. Chung, and L. Lu, "A random graph model for massive graphs," *Proceedings of the thirty-second annual ACM symposium on Theory of computing - STOC '00*, 2000. [Online]. Available: http://dx.doi.org/10.1145/335305.335326

[14] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, p. 41–42, May 2001. [Online]. Available: http://dx.doi.org/10.1038/35075138

[15] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, p. 651–654, Oct 2000. [Online]. Available: http://dx.doi.org/10.1038/35036627

[16] J. P. K. Doye, "Network topology of a potential energy landscape: A static scale-free network," *Physical Review Letters*, vol. 88, no. 23, May 2002. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.88.238701

[17] I. Liiv, A. Vedeshin, and E. Täks, "Visualization and structure analysis of legislative acts: A case study on the law of obligations," in *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ser. ICAIL '07. New York, NY, USA: ACM, 2007, pp. 189–190. [Online]. Available: http://doi.acm.org/10.1145/1276318.1276353

[18] D. Bourcier and P. Mazzega, "Codification, law article and graphs," in *Proceedings of the 2007 Conference on Legal Knowledge and Information Systems: JURIX 2007: The Twentieth Annual Conference*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2007, pp. 29–38. [Online]. Available: http://dl.acm.org/citation.cfm?id=1565610.1565618

[19] E. Täks, L. Vohandu, A. Lohk, and I. Liiv, *An experiment to find a deep structure of Estonian legislation*, 12 2011, vol. 235, pp. 93 – 102.

[20] A. Kütt, "Measuring complexity of legislation. a systems engineering approach," June 2017, paper presented at the International Conference on Artificial Intelligence and Law, London, UK.

[21] K. Sinha, "Structural complexity and its implications for design of cyber-physical systems," Ph.D. dissertation, 2014. [Online]. Available: http://dspace.mit.edu/handle/1721.1/89871

[22] A. N. Kolmogorov, "Three approaches to the quantitative definition of information*," *International Journal of Computer Mathematics*, vol. 2, no. 1-4, p. 157–168, Jan 1968. [Online]. Available: http://dx.doi.org/10.1080/00207166808803030

[23] World Bank Group. (2017, November) Doing Business - Measuring Business Regulations. [Online]. Available: http://www.doingbusiness.org

[24] ——. (2018) Methodology for Starting a Business - Doing Business - World Bank Group. [Online]. Available: https://web.archive.org/web/20180905211845/http://www.doingbusiness.org/methodology/starting-a-business

[25] ——. (2017, November) Historical Data - Doing Business. [Online]. Available: http://www.doingbusiness.org/Custom-Query

[26] ——. (2017, November) Law Library - Doing Business. [Online]. Available: http://www.doingbusiness.org/law-library

[27] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2018. [Online]. Available: https://www.R-project.org/

[28] L. A. Adamic, B. A. Huberman, A.-L. Barabási, R. Albert, H. Jeong, and G. Bianconi, "Power-law distribution of the world wide web," *Science*, vol. 287, no. 5461, p. 2115, Mar 2000. [Online]. Available: http://dx.doi.org/10.1126/science.287.5461.2115a

[29] S. N. Dorogovtsev and J. F. F. Mendes, "Evolution of networks," *Advances in Physics*, vol. 51, no. 4, p. 1079–1187, Jun 2002. [Online]. Available: http://dx.doi.org/10.1080/00018730110112519