Codification, Law Article and Graphs

Danièle BOURCIER a,1 and Pierre MAZZEGA b
a CERSA CNRS – Univ. Paris II
b LMTG CNRS - IRD - Univ. Toulouse III

Abstract. Policy makers, judges, citizens all deplore the increasing complexity of law. Codification is traditionally one of the best ways to sustain the dynamic and necessary evolution of laws. In this paper we analyze the French environmental code (2000) to understand how the drafters have organized the previous laws scattered in various fields of law into the structure of a hierarchical table of contents. Relying on graph representation, we observe this architecture through the various levels of its organization and connections with other legal corpuses. Among the lot of information that this representation brings, we will note the new function of the article of law in the pyramid of laws. The article is no more one axiom, one rule as in the Civil Code (1804): it becomes an articulation between both an upper and a lower (intra-article) hierarchy, embedded in a network of links. So doing we also find some invariant distributions – like the Pareto-Zipf one – that shape various statistical distributions of the vertices and edges of the legal graph. We will conclude by proposing new comparative perspectives between legal sources and between Codes.

Keywords. Codification, law article, graph representation, tree, network, Pareto-Zipf distribution

1. Introduction

The growth of national and international legal corpus and the rapidity of changes become more and more complex to manage for administrators. Impacts of this complexity on the intricacy of both legal hierarchies and legal contents are considered as causing other unwanted by-effects on the intelligibility of the laws for citizens. In December 2005, the French Constitutional Court has advocated that an article of a law can be considered as unconstitutional if its intelligibility and accessibility are surpassing a “reasonable complexity”1.

One of the solutions adopted by France since 1989 and now by the European Union is to gather and “codify” laws by fields in a global project. It means that in a first stage, articles of laws and decrees are reorganized following a table of content representing the structure of this field. By representing the internal and external links put by the group in charge of the codification (hereafter the “codificator”) we can follow the various structures and networks which connect paragraphs, articles, chapters, titles in the same code and with the other 60 codes for the whole French law. We consider that 75 % of the legal corpus is already “codified”. This intensive codification process allows to clear up and to organize the legal knowledge. However, codification

can only stabilize a set of laws and decrees for a limited time. The evolution of law constrains the structures of codes to evolve. In a previous paper, we drew guidelines in order to observe the complexity of law and to develop both a "structure-based" measure and a "content-based" measure of objective complexity in a legal system. In this paper, we will stress on the structure of law when it is reorganized in one code and between codes and how the graphs of this structure can evolve in time. Thus we show how we can use graphs to represent and calculate some measures of the complexity of law from the codification of the field of environment and to put new lights on legal systems and on their evolution dynamics. We search for a provisional understanding of legal complexity. For example some of the assumptions underlying these measures of legal complexity are explicitly stated and discussed. We obtain some legal network that can be characterized by different kinds of mathematical measures. These measures are summarizing some various properties of the network structure. Though it is not the purpose of this paper to develop in detail such measures, several potential research issues are clearly identified.

We will develop hereafter some graphs models giving more precise measures of what is a structure-based complexity that the codificator and citizens have to cope with.

2. Legal Knowledge, Codification and Law Article

Legal knowledge is represented by interconnected networks of rules (cognitive content) and texts (material documents). Most of the legal documents are numbered in articles. The difference between rules and texts is the following: texts are always represented in a documentary hierarchy with a name and a position. It can thus be identified and quoted in a corpus of full text. On the contrary, a legal rule is not directly represented in a chain of character. For example, several rules can be localized in the same text and their substance can be interpreted differently. In this paper, we will consider only texts and structures.

Legal knowledge has a second characteristic: even well-structured it is generally considered as complex. We propose to characterize objective complexity in legal domains by: a) a large number of interacting components; b) a partial knowledge of the (dynamical) links between components; c) a limited predictability of the system evolution and/or outputs; d) a dependency of the system properties when observed at different levels of its organization. It should be noted that we intentionally do not separate the two main sources of our limited understanding of a complex system: a) its organization or structure intricacy and its intrinsic instability; b) our limited knowledge of the components ties and evolution rules.

We will take for a first experiment the legislative part of Environmental Code (LEC for short). The LEC has been voted by the Parliament in 2003. The 6 main books have been completed later on by book VII dedicated to environmental protection in

---

2 Between 01/07/2005 and 31-12-2006, the number of articles has evolved from 2164 to 2604 (cumulated variation: 20 %).
Antarctic. In 1266 articles, this legislative part gathers the rules of 39 laws previously scattered and those coming from texts voted after the first adoption. A few HTML tags are available in the digital file of the Code, those for example marking the beginning and end of each article. However, most items searched for in this study for the construction of the underlying graph and its characterizations are not marked. So we are developing a FORTRAN computer code that allows to find, to order, to classify and to count these items of interest (hierarchical levels, objects in classes, references, positions, etc.).

Our approach is strictly empirical. The version of the LEC mentioned is our only source of information for the graph analysis. Any text quoted in the text is within our corpus if it belongs to the LEC or is considered as an external text otherwise. This last group of texts is said to be adjacent to the graph: they are connected to the LEC graph but not part of it. The sources of these adjacent texts will be seldom noted when required for the interpretation of the results. The choice of a given corpus has two main consequences: 1) it uniquely defines the graph associated to its structure: the objects to be identified as vertices, the links that will give the edges, and the whole perimeter of the graph; 2) though some latitude exists in choosing a resolution in the analysis (see below), only objects actually present in the corpus can appear in the graph. A different graph will be associated to an antecedent or future version of the LEC, or of course to another code. From a statistical point of view this means that the LEC at hand is seen as one realization of a kind of complex, time dependent, stochastic process. Conversely only an interpretation of the results within some legal language and theory can provide pertinent meanings to the raw statistical outcomes.

3. The Underlying Graphs: Trees and Networks

As will appear here below, the aims and limits of the codification process easily find a translation in the language of graph representation. The Environmental Code is divided in a legislative and a regulatory part. Each part is divided in books, each book in titles and titles in chapters themselves composed of sections, sub-sections and paragraphs. These eight classes of objects shape the canonical hierarchical organization of the Code. The next lower level is the class of articles. However we also observe some hierarchical organization within articles: paragraphs with roman numbering can gather sub-paragraphs with Arabic numbering. Some articles also have one or several nota. These two kinds of intra-article paragraphs and nota will be counted together in the class of intra-article objects for numerical estimates. Nota are not in a hierarchical dependence under intra-article paragraphs. So we end with a hierarchy consisting of 11 levels (or classes of objects) from the top “code” to the class of Arabic paragraphs.

The formal representation of this hierarchical organization of the LEC (or of the whole Code as well) is a tree-like graph. However we also find in the LEC references and quotations to legal texts: some of them belong to the LEC, some other are external to it. In other words some network structure is superimposed to the tree structure of the LEC, and the LEC is connected to other external tree-like and network structures: other

---

3 In 2005, the regulatory part (décrets, arrêtés) has been published but we will not take it into account in our experimental corpus.
codes and “nebulae” of other texts issued from a diversity of internal, European and International sources of Law. Let us formalize such an organization.

A graph \( G=(V,E) \) can be defined as a pair composed of a set of vertices \( V \) and a family of edges \( E \). An edge \( e_i \) starts at \( v_{i-1} \) and ends at \( v_i \) if a relation \( R \) exist between these vertices. In the following we will distinguish two kinds of relations between vertices, say the influence relation \( R_I \) and the selection relation \( R_S \). \( R_I \) and \( R_S \) will induce two sub-graphs respectively noted \( G/V_I \) and \( G/V_S \), the set \( V_I \) (resp. \( V_S \)) gathering those vertices of \( V \) entering the influence (resp. selection) relationship only. We can also precise the sub-graph we are working with by putting some informative exponent in the name; for example, the graph \( G/V_L^{13} \) is the sub-graph associated to the title II of book I of the LEC drawn in Figure 1. If the number of edges between any pair \((v_i, v_j)\) cannot exceed the integer value \( p \), \( G \) is said to be a \( p \)-graph. In a 1-graph, edges being distinct from each others, \( E \) is a set and not a family.

3.1. Vertices

Scanning the text of the LEC we see that objects of any class in the hierarchy can be explicitly quoted in an article. This point is illustrated by the following kind of references: to a code (art. L121-11) or to a part of a code (L428.1), to a book or a title (L213-11-5, L.151-1), to a chapter (L215-14), a section or a sub-section (L218-21), to a paragraph (L218-3). We also find references to other articles or subdivision of articles (intra-article roman or Arabic paragraphs). Some references are even done to indented lines (L.214-17), to particular sentences (L515-20) or words (L653-1). These last three kinds of quotations are subsumed to a single class of quotations in this analysis. These empirical observations imply that any object of any class can be involved in an influence or selection relation with another object and as such is a vertex of the graph. It should be noted that not all levels are necessarily represented by an object of its class in the hierarchy. For example in Fig. 1 no levels are represented between Chapter II and the subordinated articles L132-1 and L132-2.

3.2. Edges

Two kinds of links or edges – must be considered: those corresponding to self or cross quotations, and those induced by the codification process.

3.2.1. Influence-type Edges

Links of the first kind indicate the order of succession between two nodes in the proper tree-like structure. This structure results from the cognitive organization of the legal matter in the given corpus. Law articles indeed are gathered together under the proper chapter heading and their ordering reflects a rational organization of their legal content. This kind of link expresses a “legal influence” between nodes in the sense that they show a dependency oriented from links to nodes. The start and end vertices of a given edge do not necessarily belong to two successive class levels. In our analysis we will also consider intra-article levels given by roman and Arabic paragraphs (we will also count nota as a kind of subdivision).
3.2.2. Selection-type Edges

Quotations are the ties of the second kind. A law article refers for example to another law article embedded in a different legal context, because of its generic content with regard to the quoting article. For example, an article ruling the management of some body of water might refer to some article ruling the responsibilities attached to a territorial authority. This kind of ties expresses a “legal selection” oriented from nodes to ties. The starting vertex is of course always an object of the article class and we shall only look if it is a simple or a complex (subdivided) article. The end vertex can be an object of any class, belonging to the Environmental Code, to another Code, to a legal text of Internal Law or to the corpus of European or International Law.

Figure 1: Sub-graph G/V_{L13} associated to title III of book I of the Legislative part of the Environmental Code, with a resolution down to the article level. Levels (classes of objects) are indicated in the right column (from title T to article A). The smaller spheres represent vertices belonging to the LEC. The larger vertices (bottom laws) are adjacent objects, external to the LEC. Influence and selection links are represented with respectively continuous and dashed edges. The identification number of an article in the LEC is formed with three digits indicating the book / title / chapter and then a sequential number. For example, the left most article is art. L131-3.

4. Statistics on the Environmental Code Graph

With our computer code, we have scanned the LEC in search for all objects (vertices) and for their hierarchical position and quotations (edges). The automatic detection of quotations is a difficult task because they are expressed with all the resources of natural language (see [6] for an other illustration of this problem). So we have defined a list of about 400 models of quotation patterns. Such a model might include wild-characters like “?” indicating the presence of any single sign, or like “*” indicating any number of any sign. In order to avoid multiple detections of the same quotation, we have arranged the patterns in a hierarchical order (detecting first the longer targets) and we tag the parts of the text matching some target previously searched. Some references must be developed like in the expression “from art. A to art. B” which obviously concerns a series of objects.

At this stage of our research, in order to try to minimize the possible errors, we have checked by hand all the quotations. At the same time we have attributed a reference class to each detected quotation (see below) allowing to build some estimates. We now present some statistical analysis of the whole LEC (section 3.1), of the tree
$G/V_I$ induced by the sole influence relation $R_I$ (section 3.2), and of the network $G/V_S$ induced by the sole selection relation $R_S$.

\[ 4.1. \text{Levels and Size of the LEC} \]

We count the number of objects in each level of the LEC. We obtain 7 books, 30 titles, 115 chapters, 200 sections, 99 sub-sections, 21 paragraphs and 1266 articles. As we are also interested in the internal structure of articles (if any) we count the paragraphs with roman or Arabic heading, and \textit{nota}. The total number of vertices is the number of vertices obtained at the resolution of the article plus the number of intra article subdivisions. We have 1785 such intra article subdivisions. The more structured or complex article (L212-1) has 20 subdivisions distributed over 2 levels. With the root of the graph ("code" object) and the ‘legislative part’ object, we have 1740 vertices at the resolution of the article, and 3525 vertices when going into the article structures.

We first observe that the number of objects in a level is not a monotonic function of the level number\footnote{One for the code level, 2 for the part level, 3 for books, 4 for titles, etc. Articles are at the 9th level and paragraphs with roman and arabic headings at the 10th and 11th levels respectively.}: there are more chapters and sections than sub-sections and paragraphs. This feature points to the cognitive function of chapters and sections: once books have separated the matter in the large, chapters and sections give the room for a connection between knowledge of the matter (here the environment) and the way the lawyer will rule this matter. Chapters and sections are central nodes in the managing of legal knowledge. The second observation is that the number of article subdivisions is of the same magnitude as the number of articles. Concerning the second observation it is likely that most of the complexity of the treated matter is hidden in highly structured articles. Another option – at least a formal option – for codification would have been to split complex articles in simple articles and avoid the creation of two deeper levels of hierarchy under the article level. However if we assume the splitting of all articles with a principle of dichotomy, the number of subdivisions would be we twice the number of articles, four time this number with two sub-levels. As shown below the nearly similar number of subdivisions and articles is explained by the existence of two article populations: a large one with simple, non subdivided articles, and a population of complex articles. This category of article casts a light on a new function of legal articles managing a combination of various sub-stratifications and networks.

\[ 4.2. \text{The Tree } G/V_I \]

One of the most impressive results of codification is to organize a nebula of articles of law into a 1-graph. Indeed, the hierarchy underlying a code corresponds to a tree where every object (vertex) is subordinated to one and only one object from an upper level. But focusing on chapters, we observe that the size (or order: number of vertices) of the sub-graphs they are heading are very unevenly distributed. Indeed, like in “thick tail” distributions, a large proportion of sub-graphs are small (few tens of vertex) and only a few sub-graphs are large. A similar behavior is observed when considering histograms at the section level, or when the size of the sub-graphs is replaced by the number of subdivisions found in the subordinated articles.
The large number of short chapters or sections fulfills the requirement of normative intelligibility recommended by several national or European committees in charge of better lawmaking (e.g. [3,4]). However the tails of the above mentioned distributions testify that this objective is difficult to reach. When going deeper in the hierarchy levels, the “thick tail” characteristic is even increasingly evident. This feature only begins to appear when plotting histograms at the chapter and section levels because the corresponding samples (115 chapters, 200 sections) are quite small (from a statistical point of view). It is obvious on Figure 2 with the histograms at the levels of articles.

4.3. The Network G/Ns

We have computed the proportion of references found in simple articles versus found in complex articles. This proportion is nearly 1 though there are almost 3 times more simple articles than complex articles and the distributions of their lengths (in number of signs) are very similar. This is a clear evidence that the network-like structure of Law is preferentially allocated to complex articles rather than to the classical simple articles. Which classes of objects (or levels) are targeted by these references? Let us precise that in a quotation like “chapter X of book Y” we only count it as a reference to a chapter, the other information showing the path to reach this particular object. We find that ~56% of the references target articles, ~17% sub-articles objects, ~5% chapters, etc. Obviously article subdivisions have a very important role in precisely connecting different part of the legal matter.

5. Legal Interpretation of the Code Graphs

5.1. Two Populations of Articles

The large peak in Fig. 2 (left) corresponds to the main population of non-subdivided, or “simple” articles. These articles obey the classical rules of legislative drafting.

![Figure 2](image)

**Figure 2:** (left) Histogram of the number of articles (y-axis) as a function of the number of their subdivisions (x-axis); (right) Same empirical distribution but on a log-plot (circles) fitted by a Pareto-Zipf like distribution (fitting line; see sect. 4.3).
A second population shapes the tail of the distribution. These articles are complex, with up to 20 subdivisions headed by an identification roman or Arabic number (plus nota objects). We think that new trends in Law are expressed in the transfer of some complexity within articles, with the need to organize the intra-article matter into this deeper hierarchy. These articles are used as envelopes, or titles that gather an intricacy of legal matter and required distinctions in the environment, ecosystems, social organization or institutions. As noted before, these articles also present a higher degree of connectivity with other legal texts (inside the Environmental Code or in other codes or French legislation, or to European or international legal corpuses) than the simple articles. This character of a high interconnection of the complex articles into a larger legal network is also confirmed when noting that there is only a weak correlation with the length of the articles (here measured with the number of sign in the text) and the number of their subdivisions. In Figure 3 (left) we see that most of the articles are less than 4000 signs long, whatever their internal structure might be. In other words, a high degree of connectivity is associated to a complex intra-article structure, not to the length of the articles.

5.2. Codification and Legal Networking: a Tense Link

The codification process organizes all the articles gathered on a subject in a hierarchical structure, a tree. But this work is done under the pressure of the external legal objects: other sources of Law are producing legal texts that must be taken into account and connected to the objects of the code. This process leads to a legal network (\(G/V_S\) being just a very small parcel of this huge graph) self-organized, dynamically evolving with time, with no possible top down control.

5.3. Some Invariant Underlying Distributions

Let us discard the fluctuations of the distribution of the articles as a function of the number of their subdivisions (Fig. 2, right) and consider its overall shape. We see on this log plot (left) that the general trend of the distribution follows quite well a Pareto-Zipf like law (regression line). This generic law is observed in a large spectrum of natural (e.g. size distribution of sand particulates), social and economical (e.g. incomes distribution) or linguistic (e.g. frequency of word uses) phenomena. It is associated with large systems of interacting entities. Though at first glance it might seem surprising to exhibit such theoretical law over the empirical distribution of articles of the LEC versus their subdivision number, we believe this matching to be one more clue for the complex processes underlying actual rule drafting and codification considered at this scale of aggregation (1266 articles).

These distributions are said invariant in the sense that they do not depend on the context of their expression. In the same spirit, the number versus size distribution of articles plotted in Fig. 4 (right) is also strongly reminiscent of some classical distributions. However, though interpretation of such findings must be considered with caution, precisely because these distributions are quite commonly found and at best only representative of a large class of dynamics (see e.g. [2]), it might be useful to compare the parameters of these distributions: a) when considering a larger legal corpus, which should support these findings; b) when analyzing different corpuses (e.g.
different codes). Such a procedure could be an operational way to characterize the structure-based complexity of legal corpuses, codified or not.

Figure 3. (left) Scattering plot: each circle corresponds to an article of the LEC which size (in # of signs) is indicated on the y-axis and # of subdivisions on the x-axis (e.g. le longest article has more than 9700 signs and 15 subdivisions); (right) histogram of the # of articles (y-axis) as a function of their size (# of signs, x-axis).

6. Legal Studies and Graphs: Discussion & Perspectives

First of all we must stress the fact that this is a preliminary study. In particular we would like to work on a larger corpus (including first the regulatory part of the Environmental Code). Then we have to rely on more robust tools for the detection of references, as those developed by De Maat et al. [5]. Then the confidence in the fitting of invariant distributions should be more rigorously performed and accompanied with confidence estimates. However this study opens many perspectives.

The simplification of Law and the improvement of the legibility of legal texts are main targets for the Governments, Parliaments and rule makers in many democratic countries. The tools we are developing are potentially useful to make surveys about the evolution of a legal corpus. Here follows an illustration. We have extracted the dates of modifications of those articles of the LEC that have been changed between 2000 and the on going year 2007. Some articles have been kept unchanged since their inclusion in the LEC. Some other articles have experienced several modifications at different dates; only these articles enter the statistics below. We see on Figure 4 (left) that the number of modified articles has steadily increased these last years (remember that 2007 is not completed yet).

The modification process is oriented towards all kinds of articles, regardless to their more or less complex internal structure (Fig. 4, right). There is no evidence of simplification or of stabilization. The Environmental Code is quite recent. The tracking of the transformations affecting a given corpus over a longer period of time, or of a few specific articles (“life paths”) would probably also be instructive about the underlying processes and tensions feeding the legal dynamics. Depending on the aim of the analysis, other information on nodes or ties can be preserved in the representation of a legal network. For example, those laws (nodes) gathered in the same code can be
endowed with a particular tag (e.g. a color) that will help the interpretation of the legal network structure. Another tag could mark the kind of liability constraints endorsed by the nodes or some group of nodes (e.g. most international legislations on the environment are not constraining for the signatory countries). Indeed tags can be bi-univocally chosen as soon as a typology of the liability constraints is set up. This approach can be applied to other legal concepts and will be explored in a future work.

Figure 4: In the horizontal plane are given the year of the modification of the LEC articles (from 2000 to the uncompleted year 2007) and the number of article subdivisions (from 0 to 20). On the vertical axis is given the count number of articles belonging to the classes defined by the horizontal coordinates (e.g. ~180 articles are within the year 2006 – 0 subdivision class on the left-hand histogram). (left) Histogram including all the articles modified; (right) Same as on the left plot but excluding the classes of zero subdivision articles.

In this paper we dealt mainly with a structure based complexity. Our aim is now to compare statistical results with other codes. We will also connect this source of complexity with the complexity of content (rules). The only way to observe the content is not the rules as we said in section 1 but ontology. In that context, a legal ontology will be defined as a set of terms connected with other terms and forming networks of semantic basins. In the future, various aspects of complexity issues in law can be usefully approached with graph representation.

References