UNVEILING REFERENCES IN LEGAL TEXTS: IMPLICIT VERSUS EXPLICIT NETWORK STRUCTURES

Jörg Landthaler[1], Bernhard Waltl[2], Florian Matthes[3]

[1]Research Associate, landthal@in.tum.de, Technical University of Munich, Department of Informatics, Software Engineering for Business Information Systems, Boltzmannstraße 3, 85748 Garching bei München, DE
[2]Research Associate, b.waltl@tum.de;
[2]Professor, matthes@in.tum.de; https://wwwmatthes.in.tum.de/

Keywords: ***Legal text analysis, citation network, network analysis, reference detection, semantic similarity measures***

***Abstract**:* *The continuously increasing amount of legal data leads to challenges in the efficient and effective handling of information contained in legal texts. Efforts made in legal informatics in combination with improvements in knowledge engineering have been proven to be valuable for legal experts. This paper describes a data science approach to unveil network structures in legal texts. Thereby, we differentiate between explicit networks induced by legal references, and implicit networks based on semantic similarities between norms. Within this paper we present analyses of the implicit and explicit networks of a concrete German law, namely the Germany Civil Code (BGB). We have developed and implemented algorithms to automatically extract references provided in the law text and to measure semantic relatedness between norms based on the number of shared nouns. The approach unveils latent structures within the law text, which are visualized as graphs. We also show a comparison of the two different emerging networks, namely implicit and explicit networks within the German Civil Code.*

1. Introduction

Networks play an important role throughout legal science and practice. The formalization of network structures has successfully been performed by prior research and is a common scientific discipline in formal research areas, such as mathematics or informatics. Within the last decades, the research done in graph theory has moved from a rather theoretical to a more and more applicable field of investigations. This transition can be observed in the field of database systems, where it has become common to persist data not only in relational databases but also in graph databases[1]. Those graph databases use the graph structure to semantically represent and query given data.

As mentioned, networks are extensively investigated in various research domains, including mathematics, informatics, physics, life sciences and also law and jurisprudence (see Chapter 4) [Bommarito and Katz, 2010]. Progresses in text mining make it more and more promising that it is possible to algorithmically unveil network structures throughout legal texts and normative legal regulations. A trend which has been foreseen by researchers decades ago [Merkl and Schweighofer, 1997]. Recent papers of relevant scientific conferences such as JURIX [Hoekstra, 2014] and ICAIL [Hoekstra, 2015] confirm those expectations. The importance of networks structures has been

[1] *Neo Technologies*, Neo4j System Properties. http://db-engines.com/de/system/Neo4j (accessed on 07.11.2015), 2015.

identified on several levels, such as to qualitatively and quantitatively analyze legal citation networks [Agnoloni and Pagallo, 2015], to unveil complexity of legal texts [Bommarito and Katz, 2010; Waltl and Matthes, 2014] and to build recommender systems in legal information databases [Winkels et. al., 2014]. As a practical application such networks could e.g. be used to support the construction of testing schemes by legal professionals. This paper is an approach to differentiate between two different network views on legal texts. Concrete research objectives are formulated in Section 2. Section 3 summarizes important prior research and related work. The paper continues with a description of the two pipelines to extract citation and semantic relatedness networks from law texts in Section 4. A case study on publicly available data, namely the German Civil Code (BGB) is given in Section 5, including a quality assessment of the extracted references. Section 6 discusses limitations of our approach and presents ideas for future work. Finally, Section 7 concludes with a summary.

2. Problem Statement

This paper aims to unveil implicit and explicit network structures in legal texts, more specifically German law texts. Network structures can be found throughout the legal system and legislation and are essential for the understanding of norms (articles). This approach narrows the broad network perspective to two basic network structures, namely the network structure induced by explicit references within law texts and that induced by semantic relatedness of norms. Within this work we will show both structures by automatically determining them using algorithms for text mining (see Section 4). This works investigates the BGB in its consolidated version from 30. April 2014 in German language. It would essentially also be possible to expand the network analysis to a larger dataset or to include court judgments, but to show and compare the two evolving network structures, the BGB with more than 2000 norms and over 150 000 words is sufficient. Moreover, the German Civil is strongly hierarchically structured into 5 books and several levels of subchapters.

#norms	#words	#nouns	#unique nouns	Ø words per norm	Ø nouns per norm
2382	153662	50517	3920	64,5	21,2

Table 2.1: Basic Metrics of the BGB (Using Python pattern.de POS tagger, Unofficial Norm Titles Excluded)

3. Related Work

The analysis of the network structure of jurisprudence has been in the focus of legal experts, computer scientists and for researchers at the point of intersection. Thereby, the different aspects of how the network structure emerges have been investigated, such as citation networks [Bommartio et. al., 2009; Agnoloni and Pagallo, 2015], topic clusters [Merkl and Schweighofer, 1997; Lu et al., 2011], recommender systems [Winkels et al., 2014; Adedjouma et. al., 2014] or categorization of documents [Schweighofer and Merkl, 1999].

[Agnoloni and Pagallo, 2015] have analyzed the citation network within ICC (Italian Consitutional Court) cases (vertices). Thereby, they have developed a parser that automatically extracts the citations (edges) of the documents (precision: 98,4%; recall: 91,7%). They were able to investigate topological properties, such as in-degree and out-degree of the nodes. [Lu and Conrad, 2011] have published results on detailed analyses of the application of a large scale soft clustering algorithm that takes content of legal documents and metadata of those documents into account to perform a topic-segmentation. The approach of clustering legal documents with neural computation has also been published years ago by [Merkl and Schweighofer, 1997]. They have already predicted the challenge of how to handle huge amount of data in legal information systems and proposed a neural computation algorithm for topic-wise clustering. [Schweighofer and Merkl, 1999] used unsupervised neural networks to support the process of categorizing legal documents. They adapted self-organizing maps to order high-dimensional statistical data extracted from the documents (keywords).

[Winkels et al., 2014] proposed a legal recommender system to support navigation in huge sets of legal documents. They stored the references between documents in a machine readable format. They created regular expressions to find linguistic patterns of citations, which they later on clustered to reconstruct the reason of the citation. They claimed that this information can be used to suggest additional relevant legislation sources to users of legislative portals. [Adedjouma et al., 2014] have used advanced natural language processing techniques to find citation patterns in legal texts, in particular in Luxembourg legislation. They proposed a two step mechanism, namely cross reference detection followed by cross reference resolution. They implemented their linguistic patterns as JAPE grammar using the GATE NLP workbench.

[Bommarito et al., 2009] use the co-occurrence frequency of lemmatized nouns to create semantic relatedness graphs of opinions and cases of the US supreme court corpus. They also extract direct citations and point out the importance of the intersection of the different network types. In contrast to this, we work purely on law texts and attempt to detect intra-document semantic relatedness between norms. Additionally, we incorporate all nouns and not only the k most frequent and also don't use percentage thresholds but naturally dropping out integer values as a parameter for the semantic relatedness graphs generation.

4. Reference Structure in Legal Texts

A classical cross-reference within a legal text connotes that two norms are related and this relationship is pointed out by the authors of the legal norm with intention, typically by means of limiting or extending definitions, rights or obligations. In another sense, this can be seen as a tool to avoid redundancy in law texts. E.g. §536b states out that "If the lessee knows of the defect when entering into the agreement, then he does not have the rights under norms 536 and 536a"[2]. Thereby, the references to §§ 536 and 536a explicitly denote, that the lessee has particular rights, that are specified in detail somewhere else. From computer science point of view, this does not only reduce redundancy but also improves maintainability and adaptability, because if the rights of the lessee as specified in §§ 536 or 536a changes, the norm §536b can be remained without adaptions. However, these references induce the network structure, with all its dependencies and connections.

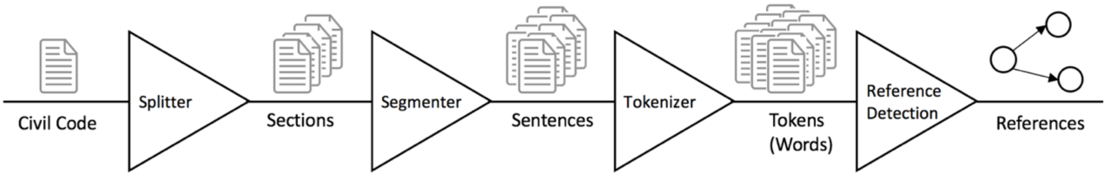

Figure 4.1: Explicit Reference Detection Pipeline Using Apache UIMA and Ruta

Figure 4.1 shows the data extraction pipeline used to determine the references in the BGB. To retrieve the referenced norms from each norm of the BGB, it was stepwise processed in a data mining framework using the Apache UIMA[3] as a base architecture. To determine the numbers of the referenced norms Ruta Scripts[4] have been used, that have been developed for this particular purpose. Moreover, using a Python script multi-references, e.g. "norms 46 to 53", have been resolved.

[2] German Civil Code BGB, http://www.gesetze-im-internet.de/englisch_bgb/englisch_bgb.html, accessed on 24.11.15
[3] Apache UIMA, https://uima.apache.org/, accessed on 24.11.15
[4] Apache UIMA Ruta, https://uima.apache.org/ruta.html, accessed on 24.11.15

Therefore, an index that maps a unique norm identifier on the norm number is necessary to include norm numbers like §50a.

Besides having an explicit reference, two norms may also be related, because they address the same real world concepts. Of course, this is a different kind of relatedness, in particular that it is implicit and much subtler and fragile.

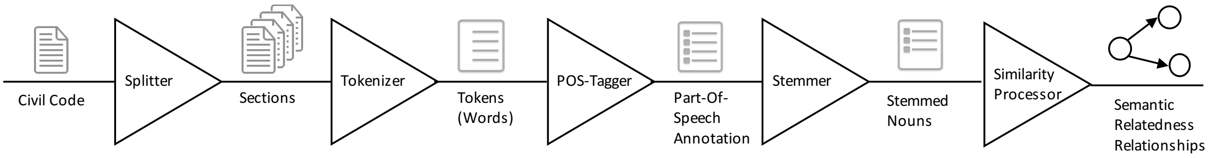

Figure 4.2: Semantic Relatedness Detection Pipeline Using Python

In linguistics it is common that nouns qualify at most of all word types to describe real world concepts – or concepts in general, e.g. "place of birth" or "dignity (of human beings)". Hence we assume that two norms that share many nouns are "related", because they address equal or at least semantically related concepts. Various semantic relatedness measures have been proposed in science, e.g. cosine similarity in the vector space model, [Salton et. al, 1975].

Similar to [Bommarito et al., 2009] we count the number of co-occurrences of nouns between norms, but in contrast to them we count all nouns without any restrictions rather than only the most frequent nouns. This is not directly a metric in a mathematical sense, because the resulting number of equal nouns does not satisfy e.g. the "identity of indiscernibles" condition. Nevertheless, for our purposes it serves as an indicator of the semantic relatedness of two norms.

Figure 4.2 illustrates our approach to measure semantic relatedness among norms of the BGB. Using a Python pipeline, the norms of the BGB are extracted from the XML version provided on the Internet. Words are splitted on spaces, tagged by the pattern.de POS-tagger[5] (part-of-speech tagger) and only stemmed nouns (NLTK Snowball Stemmer for German Language[6]) are considered for the subsequent steps. The term frequency matrix is build containing the number of occurrences of stemmed nouns of each of the 2382 norms. Finally, the vectors representing the norms are compared to each other by counting the co-occurrence of nouns. Our graphs are created by using an integer number of co-occurring nouns and we consider two norms semantically related, if the total number of co-occurring nouns is larger than a manually chosen integer threshold.

5. Case Study: BGB

The BGB is a rather huge law text compared to other German Federal Laws and the application of the references extraction pipeline depicted in Figure 4.1 yields a total of 2991 references between norms of the BGB (including 853 additionally resolved multi-references). Norms and their references can be visualized as a directed graph, where each node represents a norm and each edge represents a direct reference. We plotted the resulting references network in Figure 5.1. Colors indicate the book where the norm is hierarchically embedded in. It is known that book 1 of the BGB provides basic rules for the remaining books (clamp technique, dt. Klammertechnik). This is reflected in the resulting graph, where the norms of books 2-5 build clusters, while the blue nodes of book 1 can be seen as a central "glue" between the other books.

Moreover, graph representations allow for the application of graph algorithms, e.g. to find cycles in directed graphs. Figure 5.3 a) depicts a simple detected cycle: §81 and §83 reference each other.

[5] Pattern.de POS-Tagger, http://www.clips.ua.ac.be/pages/pattern-de, accessed on 04.12.2015
[6] NLTK 3.0 Documentation, Snowball Stemmer, http://www.nltk.org/api/nltk.stem.html, accessed on 04.12.2015

The quality of the results of the application of the references extraction pipeline has been manually determined on the first 510 norms of the BGB. Under the assumption that references are evenly distributed on average (which is supported by the ratio of the detected references) this leads to a confidence of about 98%. The precision is calculated as true positives / total predicted references ≈ 97% and the recall is calculated as true positives / total actual references ≈ 97%, see also Table 5.1.

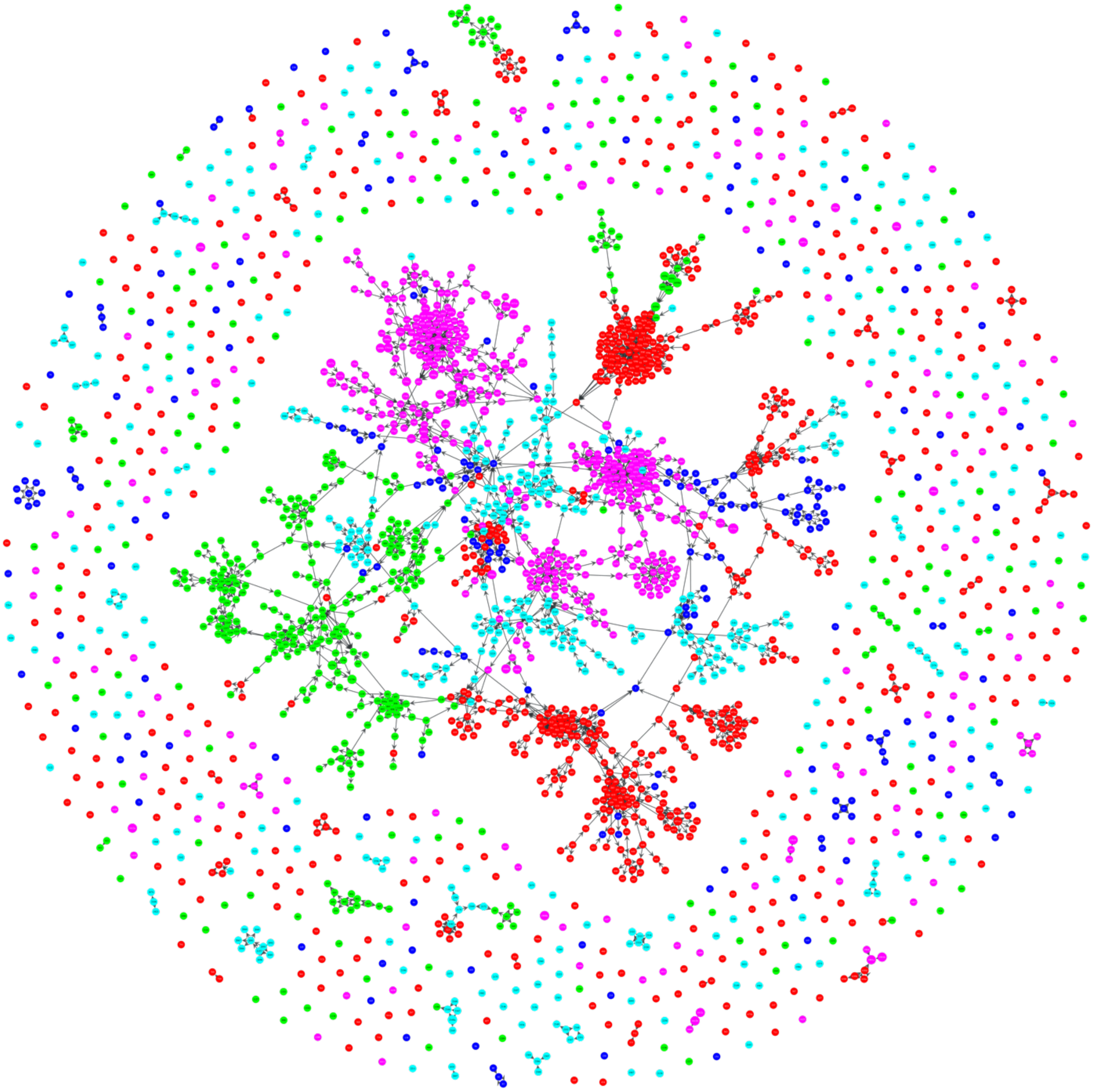

Figure 5.1: Explicit References Graph of the BGB (Blue: Book 1, Red: Book 2, Green: Book 3, Pink: Book 4, Cyan: Book 5), Seen Best in Color

Next, the results of the application of the semantic relatedness pipeline illustrated in Figure 4.2 depend upon the threshold parameter N, the integer number of total co-occurring nouns. Table 5.2 provides the results of a parameter study of the threshold parameter N. The maximum number of stemmed equal nouns between two norms is 22 (excluding the comparison of identical norms), but the results for thresholds above 12 have been omitted, because the number of covered norms drops rapidly. Again, the result can be displayed as a graph. However, in contrast to the citation network graph, these graphs are undirected. Nodes represent norms and edge represents that two norms share N or more stemmed nouns in total. Figure 5.2 displays selected graphs. Note that graphs with a larger threshold N are a sub-graph of graphs with smaller N. The graph shown with threshold N=5 includes the isolated norms as isolated nodes in order to illustrate the ratio of connected to isolated norms visually. The remaining graphs do not include isolated norms, but show more details of the structure of the graphs.

		Prediction Outcome	
		Reference	**No Reference**
Actual Outcome	**Reference**	465	16
	No Reference	15	-

Precision $\dfrac{465}{465+15} \approx \mathbf{97}\,\%$

Recall $\dfrac{465}{465+16} \approx \mathbf{97}\,\%$

Table 5.1: Quality Assessment of the Reference Extraction Pipeline on the BGB

As expected, the number of isolated norms increases with a larger threshold N, while the number of edges (and hence also the degree of nodes and the number of connected norms) drops. The size of the largest connected component drops with larger threshold values N, too.

N	CN	IN	LC	TE	MO	CE	P
0	2382	0	2382	2835771	2381	2984	0.1
1	2379	3	2379	586726	1706	2255	0.4
2	2312	70	2312	143516	952	1473	1
3	2073	309	2064	36396	516	926	2.5
4	1597	785	1530	10650	274	584	5.5
5	1099	1283	972	3862	151	370	9.6
6	747	1635	524	1596	80	238	15
7	501	1881	280	792	49	157	20
8	331	2051	140	387	30	89	23
9	213	2169	77	209	19	58	28
10	140	2242	52	123	13	46	37
11	107	2275	35	81	8	33	41
12	77	2305	22	55	7	19	35

Table 5.2: Parameter Study on Larger or Equal Number of Stemmed Nouns N Between Norms in the BGB (CN = #Connected Norms, IN = #Isolated Norms, LC = #Norms in Largest Component, TE = Total # Edges, MO = Maximum Degree of Norms, CE = #Common Edges, i.e. Edges Existing in Cross-References Graph and Semantic Relatedness Graph, P = Percentage of Common Edges to Total Edges ≈ CE/TE)

[Bommarito et al., 2009] already suggested that the intersection of the references graph and the semantic relatedness graphs might comprise clusters of topical domains. To examine this intersection, we count the number of edges existing in both: the references graph and the semantic relatedness graphs. The P column of Table 5.2 shows the percentage of common edges compared to the total number of edges respective semantic relatedness graph. The number of edges in the references graph is constant, but for larger thresholds N the percentage of common edges increases even for semantic relatedness graphs with fewer total edges than edges in the citation network graph. This supports our assumption that norms sharing a larger number of stemmed nouns are indeed semantically related stronger, under the assumption that referenced norms are often related semantically stronger than norms that do not have a reference in either direction.

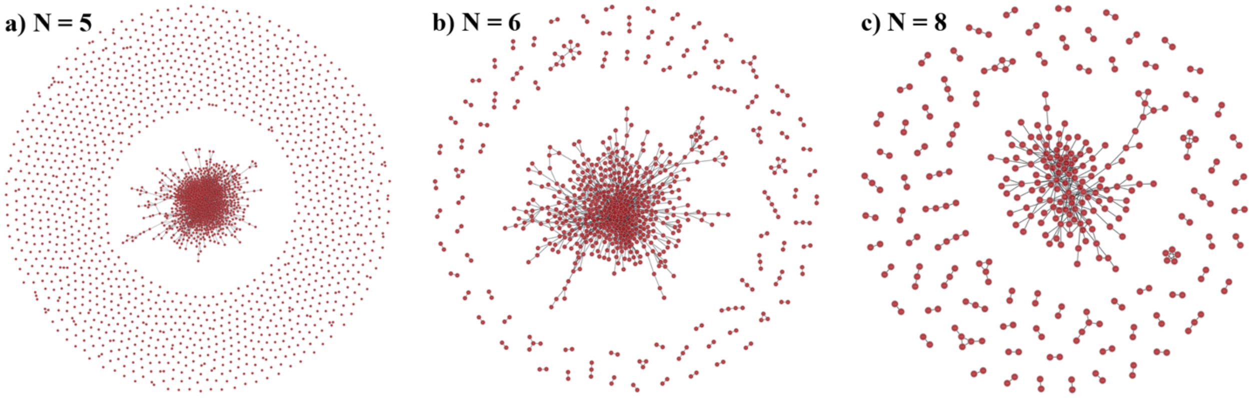

Figure 5.2: Implicit Relatedness Network Graphs for Greater or Equal Number of Nouns N, a) Including Isolated Norms, b) and c) Without Isolated Norms

Figures 5.3 b) and c) show selected details of Figure 5.2 c) (N=8): From a graph theory point of view it is interesting that the norms shown in b) form a clique, all covering the topic of "limitation of claims". Many edges connect neighboring norms, e.g. §§ 675p,i,k shown in c). But, some edges connect norms from books 4-5 to norms of book 1, where no direct reference exists. It could be interesting to investigate § 2196 ("Impossibility of fulfillment") to get a hint that there is a general norm § 527 ("Non-fulfilment of the condition"). A similar relation exists between § 2123 ("Economic plan") and § 1038 ("Economic plan for forests and mines"). However, we leave the judgment of these results to the legal experts.

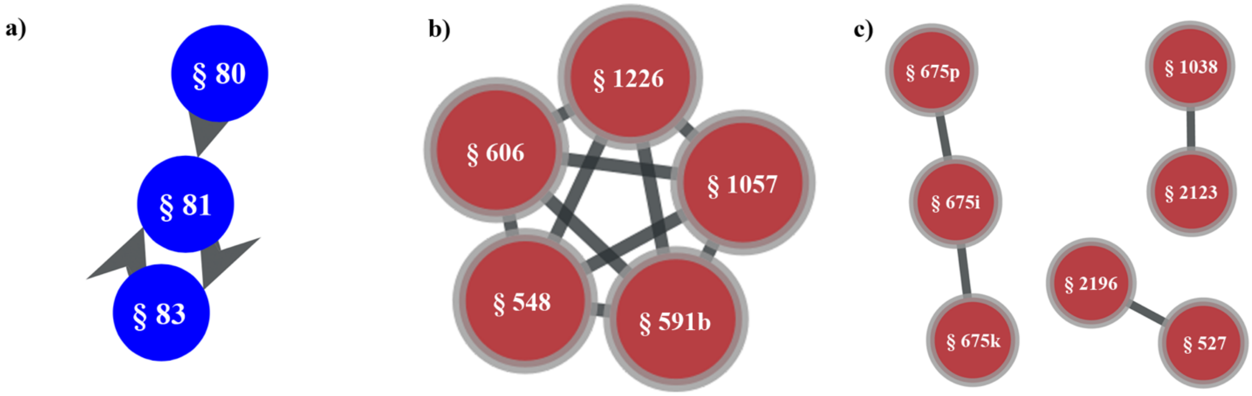

Figure 5.3: Selected Sub-graphs: a) A Cycle in the Explicit References Network Graph, b,c) Sub-graphs of the Implicit Relatedness Network Graphs for Greater or Equal Number of Nouns N=8

6. Limitations and Future Work

The references extraction is based on linguistics rules (patterns) using the fact that explicit references to norms in the BGB are prefixed by a paragraph sign. Certain references need to be resolved, e.g. "norm 46 to 53". The hardest part is to distinguish between internal references and references to external laws. This distinction needs to be incorporated manually using a complex script language, e.g., Apace Ruta. However, these pattern definitions can be applied to other laws, but may require adaption to specific conventions. The granularity of our pipelines is restricted to norm level.

Semantic relatedness is based on the number of stemmed equal nouns shared among norms. Therefore, longer norms are much more likely to be related to other norms due to their larger vocabulary (more nouns). This could be relaxed by weighting nouns with respect to their occurrence frequency and other measures, e.g. the entropy-based information gain or ontology based measures. Consequently, this effects the threshold parameter N. One need to express the relatedness as float value, and no longer as integer representing the number of total equal nouns shared among norms. However, this threshold parameter needs to be chosen manually in our pipeline, too.

A promising idea is the exploration of common sub-graphs among the references and the semantic relatedness graphs. Moreover, it is possible to exploit the hierarchical structure of the norms to find topical or other notions of relatedness among norms of law texts. Finally, the references and semantic relatedness networks could be extended to court decisions or other types of legal text documents.

7. Summary

This paper presents a comprehensive and data-intensive approach to analyze the network structure of the German Civil Code. It differentiates between two network structures that emerge in the law text, namely the explicit network codified by references and the implicit network, which arises through the semantic relatedness. The analysis is done via two data-analysis pipelines extracting direct references and semantic relatedness ("bag-of-words", "bag-of-nouns") among different norms within a law text. The two resulting networks can be displayed as graphs. The highly accurate pipelines are applied to the German Civil Code and selected resulting graphs are depicted. Besides the visualizations of the two networks, which serve as a starting point for analysis and exploration of the data and analytical information is provided, e.g. important graph metrics. We provide a quality assessment of the citation extraction algorithm and analyze the intersecting edges of both networks.

8. References

Adedjouma, Morayo/Sabetzadeh, Mehrdad/Briand, Lionel C., Automated detection and resolution of legal cross references: Approach and a study of Luxembourg's legislation. In: Requirements Engineering Conference (RE), 2014 IEEE 22nd International, 2014, p. 63–72.

Agnoloni, Tommaso/Pagallo, Ugo, The case law of the Italian constitutional court, its power laws, and the web of scholarly opinions. In: Katie Atkinson und Ted Sichelman (Hg.): the 15th International Conference, San Diego, California 2014, p. 151–155.

Atkinson, Katie, ICAIL 2015: Proceedings of the 15th International Conference on Artificial Intelligence and Law (2015). New York, NY, USA, 2015.

Bommarito II, Michael J./Katz, Daniel M., A mathematical approach to the study of the United States Code. In: Physica A: Statistical Mechanics and its Applications 389 (19), 2010, p. 4195–4200.

Bommarito II, Michael J./Katz, Daniel/Zelner, Jon, Law as a seamless web? comparison of various network representations of the United States Supreme Court corpus (1791-2005). In: Proceedings of the 12th International Conference on Artificial Intelligence and Law. Barcelona, Spain 2009, p. 234–235.

Hoekstra, Rinke, Legal Knowledge and Information Systems: JURIX 2014: The Twenty-Seventh Annual Conference: IOS Press (Frontiers in artificial intelligence and applications), 2014.

Merkl, Dieter/Schweighofer, Erich, En route to data mining in legal text corpora: Clustering, neural computation, and international treaties. In: Database and Expert Systems Applications, 1997. Proceedings., Eighth International Workshop on. IEEE, 1997, p. 465–470.

Lu, Qiang/Conrad, Jack G./Al-Kofahi, Khalid/Keenan, William, Legal document clustering with built-in topic segmentation. In: Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011, p. 383–392.

Salton, Gerard M/Wong, Andrew/Yang, Chungshu, A vector space model for automatic indexing. Commun. ACM 18, 1975, p. 613-620, 1975.

Schweighofer, Erich/Merkl, Dieter, A learning technique for legal document analysis. In: Jon Bing, Andrew J. I. Jones und Thomas F. Gordon (Hg.): the seventh international conference. Oslo, Norway, 1999, p. 156–163.

Waltl, Bernhard/Matthes, Florian, Towards Measures of Complexity: Applying Structural and Linguistic Metrics to German Laws. In: Jurix 2014: Legal Knowledge and Information Systems, 2014.

Winkels, Radboud/Boer, Alexander/Vredebregt, Bart/van Someren, Alexander, Towards a Legal Recommender System. In: Frontiers in Artificial Intelligence, Volume 271: Legal Knowledge and Information Systems, 2014, p. 169–178.