# Chasing the Invisible in the Grammar of Repetitions: A Network Analysis Approach to Fiscal State Aids

Galileo Sartor[1,*,†], Piera Santin[2,†] and Luigi Di Caro[1,†]

[1]University of Torino, Torino, Italy
[2]University of Bologna, Bologna, Italy

Abstract
The significance of interpretation principles seems to be inexorably tied to the moments in time when they are recurrently referenced by scholars or judges. Moreover, these sequences of references may undergo substantial shifts in meaning or contextual usage over time. Based on this phenomenon, our work proposes a network analysis approach to identify and locate such pivotal points. Specifically, we begin by extracting and mapping citations in judicial rulings, focusing on the specific context of fiscal state aids in the case-law of the Court of Justice of the European Union. We then demonstrate how applying network analysis to these citations can serve as a valuable tool for enriching the legal study of CJEU case-law. In detail, we focused on the network of precedents as cited by the Court to verify how the case-law develops new interpretative principles and contributes to the creation of a legal framework for European discipline of fiscal State aids. To retrieve the necessary information on precedent references within a judgment, we utilized the XML representation accessible on the EUR-Lex platform. We then employed regular expressions to parse the text and guarantee the precise and complete extraction of citations. Our research highlights how automated analysis of citation networks can offer valuable resources to supplement conventional legal methodologies.

**Keywords**
network analysis, legal knowledge extraction, citation networks, text similarity, CJEU

## 1. Introduction

It is common knowledge that, for European tax law and for European law more in general, the case-law of the Court of Justice of the European Union (CJEU) plays a key role. To some extent, case-law contributes to defining and realising the single market, which is the ultimate objective of the Treaties. However, from the point of view of the system of sources, they remain a *unicum*, an undefined legal object which distinguishes the role of EU jurisprudence from the traditions of the member states, both in common law and civil law.

The significance of interpretive principles has become increasingly apparent over time, as the Court of Justice frequently employs them as pivotal points in its exegesis of European law. While literature has emphasized the importance of these principles, it has yet to explore how they achieved such prominence. To investigate the development of citations and ensure the verbatim nature of references to precedents, we opted to integrate quantitative techniques into our analysis[1].

Our analysis centers on citations to case-law within CJEU judgments in the domain of Fiscal State Aid (excluding legislative references). This approach enables us to evaluate how the CJEU employs case-law citations to bolster and shape judicial decision-making. All the extracted data, the code and the output of the developed system is publicly available at https://github.com/LyzardKing/citation_extraction.

## 2. Background

In a new case, the Court frequently cites precedents by quoting a specific paragraph (or a few) that encapsulate a significant concept or principle. This implies that citations are typically not intended to refer to the entire judgment, but rather to the specific paragraph in question. For this reason it is common to find citations to case

The structure of the available versions of judgements, and other documents of the CJEU adheres to a set of openly available rules, adopted by the Court of Justice of the European Union.

The method of citing the case-law in particular combines the ECLI with the usual name of the decision and the case number in the register. It has gradually been brought into use by each EU Court/Tribunal since the first half of 2014, and was harmonised as between the Courts of the European Union in 2016. The reference comprises several elements, including the type of decision (i.e., judgment or order), the complete date of the

*Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023), June 23, 2023, Braga, Portugal.*
*Corresponding author.
†All authors contributed equally to this research.
galileo.sartor@unito.it (G. Sartor); piera.santin2@unibo.it (P. Santin); luigi.dicaro@unito.it (L. D. Caro)

[1]For the title of our paper we have been inspired by Jacques Derrida, *Writing and Difference*, University of Chicago Press, 1978

decision, the customary name of the case, the case number in the register (e.g., for the CJEU C-nn/aaaa), the ECLI, and the cited paragraph.

This fixed citation style has many advantages, such as:

- it improves the accessibility of judicial decisions by ensuring that references to case-law consistently include all necessary information to unambiguously identify the decision in question. Specifically, each reference includes all constituent elements, which are reiterated every time the reference is made;
- it provides greater linguistic neutrality since the format of the citation is largely identical in all languages and thus contains fewer elements to be translated;
- it facilitates the automatic insertion of hyperlinks on the ECLI of the cited decision and to the relevant paragraph.

It is important to underline that the case number in the register identifies all the documents referable to a specific procedure, which means judgement of any grade, opinion of the Advocate General, order, request for a preliminary ruling. Vice versa, the ECLI identifier refers to a single specific document, and allows for the identification of the cited judgement. That means that different ECLI may refer to different documents related to a single case.

We opted to adopt the Court's own method of citing precedents, focusing on the cited paragraphs rather than the judgment as a whole in our network analysis. This choice was justified also by the fact that, as mentioned, the cited paragraphs may pertain to a judgment on a different topic from the one of the citing document. One such example can be found in judgement Case C-322/09 P NDSHT, a judgement of appeal in field of fiscal State aid, where paragraph 41 cites case C-229/05 P PKK and KNK v Council [2007] ECR I-439, paragraph 66. The second case is not on fiscal State aid, but instead deals with *restrictive measures directed against certain persons and entities with a view to combating terrorism.*

Such connection may appear perplexing if we consider it in the network of judgements on state aid, but is instead perfectly reasonable if we look at the concepts expressed in the two paragraphs. In particular both concern the issue of the arguments that an appellant is allowed to put forward, a fact which is equally relevant in the two judgements.

Furthermore, the citation of paragraphs, regardless of the topic of the judgements, concurs in giving them an autonomous value as interpretative principle. As interpretative principles, they may be used in the CJEU argumentation as part of the European interpretative framework, almost comparable to a legal rule.

At this stage, we choose to work only on explicit citations, without looking for the implicit ones. Hence, our objective is to examine the connections that judges aim to establish when constructing their argumentation framework. It is worth noting that, particularly for widely recognized principles, a judgment citation involves a deliberate selection from numerous precedents, oftentimes identifying both the most and least recent. This potential limit has been overcome with the linking of direct and indirect citations (see subsection 4.3). We assume, and partially demonstrate, that the judgements that the Court decided to refer to are generally the oldest ones, until the reaching of a "canonization" of the interpretative principle, that from then on is the cited case (see section 4.3.2).

We chose to focus on the field of fiscal State aid, since this topic is representative of the creative role played by the CJEU (particularly in the field of fiscal State aids), and because it is a small enough field that the connections seemed more verifiable in an initial assessment. By concentrating on a particular field, we are afforded the opportunity to assess the effects of citation network analysis utilizing a methodology that diverges from previous investigations. In particular, we tried to merge the methodological and meta-argumentative paths, connected with the use of precedents by European Judges, with an analysis of the actual impact on specific legal issues.

## 3. Related work

We are not the first to examine the citation of precedents (or other legal authorities) as a means of understanding the importance of courts, opinions, or judges (see, e.g., [1]; [2]; [3]; Sirico 2000; [4]). We thus follow in a long tradition of examining legal citations, but we show that recent advances in the methodology of network analysis lead to more nuanced and precise measure of the relevance of a case for the network of law, following the example of [5].

Moreover, all the above mentioned studies have been developed with regards to common law systems, which rely on the concept of binding precedent rather than of positive rule. According to this, the importance of precedents is significant in such systems and many studies have focused on it since a long time.

The approach is slightly different when it comes to the CJEU which is not part of a common law system [2].

For this reason, there is only a reduced number of studies of citation networks specifically dedicated to CJEU case-law, and all of them underline difficulties in work-

[2]Neither a strictly civil law one, since the European judicial system is a hybrid of the two. There is an agreement in literature on the impossibility of finding any principle about binding precedents. That fact creates an unbeatable obstacle in considering the European one as a common law system

ing with judgements. In particular we could recall the works of Derlen and Lindholm, which developed from [6] to [7]. They stress the use of network analysis mainly to evaluate degree centrality, with significant results in term of the evaluation of persuasion. Such approach has been object of an interesting criticism by [8] focused on the lack of theory about how the use of a specific type of precedent is reflected in a citation network.

According to [9] a network analysis and a similarity comparison is useless if it focuses on the full texts of CJEU court decisions, since it does not closely mirror citation behaviour and there is a substantial overlap.

Instead of ranking entire judgments, in [10] it is proposed to directly rank the cited paragraphs, in order to avoid the mentioned inaccuracy, and for the analysis to correspond more to the legal importance of the specific citation. This methodology seems to be ideal, considering the peculiar structure of the CJEU system of references.

By learning from these previous works, the dataset we will describe and use is based on single paragraphs as the main nodes, with additional information on the entire judgments.

The categorization of inconsistencies of CJEU citation policy is analyzed in-depth by [11]. The article identifies and explores three types of alteration, or mechanisms of instability: (1) the substitution of cited cases in citation strings; (2) the alternation between expressions found in settled case law and alternative expressions; and (3) the un-anchoring or detachment of legal statements from cases in which they initially appeared. The analysis illustrates how substitution leads to diverging interpretive outcomes, how alternation unsettles the normative force and the relevance of the acquis, and how un-anchoring results in a loss of knowledge.

From this perspective, references to settled cases result in complex changes to the law, which are multidirectional and lack a clear progression. We refers to the concepts of substitution and un-anchoring, trying to deal with them in the analysis of centrality and semantic similarity.

There have been, and are, multiple attempts at the automated identification and extraction of legal citations, both from case-law and legislation, with different technological means. In [12] the use of regular expressions is proposed, and we will see that it remains an important tool even in more structured datasets. Named entity recognition ([13]) and other Information Extraction methodologies ([14]) are also proposed as a more domain specific approach.

In [15] the concept of multi-dimensional citation networks is considered. This concept is also described in our methodology, and enables the structuring of a large network, considering both citing and cited paragraphs (i.e., in- and out- citations).

Going one step further from the citation extraction we have different attempts at building graph networks, mainly focusing on the US legal system, as in [16], and EU courts ([17], [18]).

In [19] the similarity and relevance of legal citations is then applied to historical cases from the Court of Friesland, and used to assess the importance of case-law citations from a historical view. The same idea is applied in [9] to judgements by the Court of Justice of the EU, to enhance the network graph with semantic and structural text analysis.

# 4. Methodology and Results

In this section, we describe the process and the tools used to build the database of case-law citations in the previously defined context.

Often, studies such as the one proposed in this paper suffer from the limited availability of structured data, requiring a long manual and preliminary work. By leveraging the cases made publicly available in declarative markup formats by the ECJ on the EurLex platform, we developed an process that allowed to directly focus on the data analysis of the research.

The objective of this work was to determine if and how an automated network analysis phase could complement the human analysis in navigating and extracting useful insights from complex citation networks [19].

## 4.1. Data corpus

The source documents on which the analysis was carried out are all available on the EurLex platform, and accessible through Cellar, the common repository of metadata and content for EurLex [3].

The first step was to verify the availability of cases in the chosen legal domain. This task was carried out by searching the EurLex and Curia databases with the following filters: "judgement" as a type of document; "CJEU" as Court (avoiding judgement of the General Court); "appeal" as procedure; "State Aid" as subject-matter; plus we added the word "tax" in the free text, in order to find cases of *fiscal* State aid. We then identified what information would need to be extracted.

Most of the cases we analyzed are available in an XML format, Formex[4], that is used to add structural information and metadata to case law. In particular, for what was necessary to the definition and analysis of citation networks, the cases contain detailed information on legislative and case-law citations, in particular with the XML

[3] as defined at https://op.europa.eu/en/web/cellar

[4] Formex describes the format for the exchange of data between the Publication Office and its contractors. In particular, it defines the logical markup for documents which are published in the different series of the Official Journal of the European Union.

tag "REF.DOC.ECR". The availability of this information made extracting references from cases much easier.

When this XML representation was not available the data extraction relied mainly on regular expressions, although this was more of an issue with older cases. Generally however the style of citation was found to be common enough to make the extraction possible without too many issues.

Having structured cases, either in XML or HTML, made it possible to extract also the cited paragraphs. These were then used to further crawl the network, by parsing the cited paragraph for other citations, as well as to compare the semantic similarity of the cited and citing paragraphs. This is a novel development in the legal citation analysis, and its usefulness and purpose will become clear in the following sections.

For the proper extraction of the paragraph number and content, the structural information contained in the XML representation is further enhanced with regular expressions. Between the two modes (XML and regex parsing) it is generally possible to extract the correct paragraph. On the one hand, methods based on regular expressions are often used in systems for text extraction and analysis [12, 16]. On the other hand, having a curated (and automated) indication of the metadata of the citations embedded in the available representation, be it XML, HTML, or other formats, makes it possible to reuse what is already available, and simplifies the data extraction phase.

From the procedural point of view, the developed tool downloads the XML representation of the case document and identifies the citations. Then, for each cited document it repeats the process recursively, building a the database of cases and citations. In the selected domain, we considered the cases from the Court of Justice, excluding those from the General Court, which in judgements of appeal such as those we are interested in, acts as a Court of First Instance.

## 4.2. Data structure

Once extracted, the citations are stored as json representations of the original XML object, as can be seen in Listing 1, with a subset of the metadata and the structure extracted. In particular, necessary information in this context is the text of the paragraph containing the citation, as will be highlighted in Section 4.3.2, and the URL pointing to the XML version on EURLex, to recursively repeat the search. Other information is available but not yet used in the network analysis, while still being captured for future developments (see 5).

The main advantage on using a json human readable representation is that it enables the legal experts involved to directly access the information, for an initial qualitative evaluation of the output, that can be used to alter the methodology efficiently.

Listing 1: Citation stored in JSON

```
{
    "ecli": "ECLI:EU:C:2021:201",
    "text": "Judgment of 16. 3. 2021
        - Case C-562/19 P Commission v
         Poland",
    "par_num": "NP0001",
    "celex": "62019CJ0562",
    "case_no": "C-562/19 P",
    "keywords": [
        ...
    ],
    "xml_url": "http://publications.
        europa.eu/resource/ecli/ECLI%3
        AEU%3AC%3A2021%3A201.ENG.fmx4.
        ECR_62019CJ0562_EN_01.xml",
    "par_text": "1By its appeal,
        [...]",
    "references": [],
    "outcome": "On those grounds,
        [...]"
}
```

The general idea is to have the complete set of citations for the cases in our domain, and to proceed with a recursive analysis only for the cited paragraphs. This enables us to analyse from a historical point of view the relevant citations, without broadening the number of citations to analyse more than strictly necessary.

From the initial list of 40 cases extracted from EurLex, we extracted a total of 1435 paragraphs, from 493 judgments. In this dataset there are 1392 relationships between paragraphs.


**Figure 1:** Representation schema.

In Figure 1 it is possible to see the definition schema of the database, with judgments (in pink), that contain paragraphs (in gray). The paragraphs refer to other paragraphs.

## 4.3. Data analysis

The database generated can be easily exported to JSON-LD, RDF, cypher, and other languages for further semantic analyses. In our case, we used the cypher lanugage, importing the data in a Neo4J database[5]. With this data

[5]https://neo4j.com

imported, it is possible to visualise the citations as a graph, with arrows going from the citing paragraph to the cited one (REFERS_TO), and back from the paragraph to the containing case (BELONGS_TO). It is then possible to query the graph as a database and extract information.

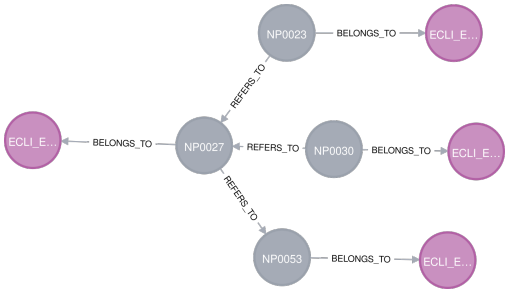


**Figure 2:** Links between cases and paragraphs

This representation allows to visualise the more frequently cited cases and paragraphs. This in itself may represent an indication of the importance of certain cases in the legal literature.

#### 4.3.1. Centrality algorithms

A more detailed analysis on the relevance of citations can be done by using centrality-based algorithms to determine the importance of nodes in the network. The metric used for this evaluation is the Degree Centrality[20], which measures the number of in/out relationships for the different nodes. In this case, we are interested in the incoming *REFERS_TO* relations between *Paragraph* nodes[6], in order to see which are cited more frequently through distinct relationships.

**Table 1**
Degree centrality values of cited paragraphs.

| ECLI | Paragraph | Citations |
|---|---|---|
| ECLI:EU:C:1994:211 | NP0059 | 12.0 |
| ECLI:EU:C:1991:161 | NP0021 | 8.0 |
| ECLI:EU:C:2001:598 | NP0041 | 6.0 |
| ECLI:EU:C:2002:506 | NP0022 | 6.0 |
| ECLI:EU:C:2011:732 | NP0087 | 5.0 |
| ECLI:EU:C:1996:64 | NP0079 | 5.0 |

This results in a list of paragraphs, sorted by the number of instances of direct citations.

The graph can be analyzed from both a *vertical* and an *horizontal* point of view, by either searching for the temporal evolution of a citation, or the frequency of citations. The vertical analysis enables the identification not only of direct citations of paragraphs (case A cites case B), but also the indirect citations (where paragraph A cites paragraph C, that in turn cites paragraph B). From a legal standpoint this allows us to recollect the historical evolution of a specific interpretative principle, as well as to have an overall view of the legal precedents referred to in a specific judgement.

To address this, it is possible to collect the citations recursively for each starting node, then summing up the distinct paths. We are interested in the distinct paths to avoid duplication, or counting the same relation more than once.

**Table 2**
Degree centrality values of cited paragraphs

| ECLI | Paragraph | Citations | |
|---|---|---|---|
| | | Direct | Indirect |
| ECLI:EU:C:1991:142 | NP0018 | 2.0 | 33 |
| ECLI:EU:C:1986:22 | NP0037 | 2.0 | 32 |
| ECLI:EU:C:1986:22 | NP0038 | 2.0 | 32 |
| ECLI:EU:C:1992:381 | NP0016 | 1.0 | 31 |
| ECLI:EU:C:1978:36 | NP0018 | 1.0 | 30 |
| ECLI:EU:C:1978:36 | NP0019 | 1.0 | 30 |

It is also possible to use other metrics or algorithms such as PageRank [21], although the use of Degree Centrality already demonstrated to return useful results.

On the basis of this representation, the vertical analysis of a specific interpretative principle may show how much the CJEU uses direct repetition and formulas (e.g. *it is settled case law* [11]), even when it modifies the precedents to create new interpretations, as can be seen in Figure 3. It is also possible to identify when the citation stops referencing directly the initial case, relying on the generalised precedent.

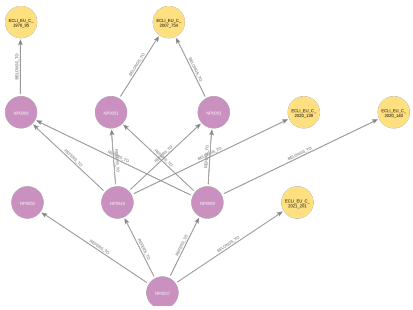

**Figure 3:** Vertical analysis

In the latter case, there is often a correspondence between the judgement which are well-known as "leading case" between scholars and nodes that has a higher number of incoming relations, as is shown with respect to 2021:201 in 3, where the citation ranking is built with the

[6]The text of the paragraph is hidden in the table, to reduce the space occupied in each row.

method described in 4.3.1, which means that we extract the total of direct and indirect (re-anchored) citations.

**Table 3**
Most cited paragraphs in ECLI:EU:C:2021:201

| Name | ECLI | Par | Cit |
|---|---|---|---|
| Tubemeuse | EU:C:1990:125 | 25 | 9 |
| Banco Ext de Espana | EU:C:1994:100 | 13 | 9 |
| Spain v Commission | EU:C:1994:325 | 20 | 9 |

One of the effects produced by repetition is what Sadl [11] classifies as un-anchoring. THis is the situation where *the foundational case, or a case from which the legal phrase originated, is omitted in the process of repetition (and over time forgotten as the original case)*, producing loss of knowledge. Similarly, un-anchoring can concern the dissociation of later legal statements and cases from the original. Often the specific meaning of the original legal statement is lost.

The most significant outcome of this experimentation is to realize how used the un-anchoring mechanism is within the selected field of fiscal State aid. The most directly cited interpretative principle is par. 59 of EU:C:1994:211, Brazzelli Lualdi et al., while if we consider the aggregate (or re-anchored) data, the most cited one is EU:1992:381, Portuguese Republic and Kingdom of Spain v Council of the European Communities. None of them involve substantial issues on State aid.

The fact that, in our dataset, the most cited paragraphs apparently do not belong to judgement in the field of State aid may have a quite simple explanation. We choose to focus on appeal judgements, since they are closer to judgements in plain litigation, since the CJEU does not merely play the role of an interpretative judge. In these cases thus there may be many procedural issues raised by the parts and discussed by the Court, allowing for many citations of a procedural nature. These kinds of issues do not depend on the substantial object of the controversy. Thus, it is reasonable that judges looked at the precedents that may be relevant for solving the procedural questions. This fact is however not explicitly stated in the citing cases, which is exactly one of the effects that un-anchoring is supposed to have.

#### 4.3.2. Semantic similarity

The final analysis carried out on the extracted information is the semantic comparison of paragraphs.

The idea behind this approach is that a vast network of citations is not easy to handle manually. Instead, an automated approach based on different similarity metrics can be used to verify the relevance of citations [9].

We are interested in a semantic-based similarity approach, to take into account not only the sequence of characters, but also the context of the contained words. In particular, we tested the Sørensen–Dice coefficient with the Cosine distance, which gave a good indication of the similarity of the cited paragraphs.

An effective example of the insights that this approach may offer in verifying substitution and un-anchoring comes from a citation chain starting from the case C:2021:201, Commission v. Poland. In detail, the three sentences are:

- **ECLI:EU:C:2021:201, par. 37:** "*As regards the fundamental freedoms of the internal market, the Court of Justice has held that, given the current state of harmonisation of EU tax law, the Member States are free to establish the system of taxation which they deem most appropriate, meaning that the application of progressive taxation falls within the discretion of each Member State. The same is true in the field of State aid*"
- **ECLI:EU:C:2020:139, par. 49:** "*However, it must be recalled that the Member States are free, given the current state of harmonisation of EU tax law, to establish the system of taxation that they deem the most appropriate, and consequently the application of progressive taxation falls within the discretion of each Member State*"
- **ECLI:EU:C:1976:95, par. 9:** "*Although this provision prevents taxes being levied on the products of other Member States which are higher than the taxes applicable to similar domestic products, it does not however restrict the freedom of each Member State to establish the system of taxation which it considers the most suitable in relation to each product*"

If we consider the Sorensen-Dice similarity index, we can calculate the distance between the first two paragraphs (those that are apparently closest), with and without the citations. This index is calculated by dividing the number of common elements in the two samples by the average number of elements.

In the first case, the similarity seems low (67.60%), compared with the actual semantic value of the text. If we remove the references to case law, the value is a higher (82.20%). This discrepancy could be attributed to what is stated in [11] as *Substitution*, described as *continuously replacing or reshuffling older and newer cases, and co-citing cases with opposing outcomes.*

This method of changing the references has a profound relevance when we focus on a specific field of law, such as fiscal State aids, as in our experiment. Another level of substitution may be the contamination with other fields of law, in which the CJEU expressed an interpretative principle for the first time.

A different approach that can be used in assessing the semantic similarity is to use pre-trained Transformer models (e.g. BERT-like) that convert input texts into

vectors (embeddings) thus capturing semantic information, and requiring less manual intervention in removing sections of text before hand. In particular, we calculated the Cosine similarity on a set of embeddings of the paragraphs[22]. To extract the embeddings we used the Sentence Transformers library, using existing pre-trained models[7].

In Figure 4 we have mapped the cosine similarity between three related paragraphs of those in Figure 3.

The comparison between semantic and non-semantic similarity score appears useful especially when the results differ, and it can be useful in the legal analysis (e.g., in identifying un-anchoring or substitution). Furthermore, discrepancies in semantic and non-semantic results may signal the presence of an anomaly, that could be interesting to in-depth analyse.

Our example regards the principle of national supremacy on tax matters, which is expressed with regards to competition and State aid (EU:C:2021:201) or to direct taxation and economic freedom (EU:C:2020:139 and EU:C:1976:95). Direct taxation is a sector of exclusive competence of Member States, so an intervention of EU institutions is possible only if it is justified by prevalent interest (like protection of one of the fundamental economic freedoms). It is very different when it comes to the field of State aids, which are regulated directly in the Treaties, with a fully European competence. In such a scenario, substitution may be used to reinforce the political statement of the Court of Justice, without any kind of mitigation due to the intersection of different subjects.

The first and second paragraphs pertain to similar field, since both cases involve progressive turnover taxes in Poland, but they are part of different legal procedures, since the first one is an appeal and the latter is a reference for a preliminary ruling. Moreover, the first one refers to the nature of State aid in turnover taxes, whereas the latter refers to freedom of establishment.

Consequently, the first consideration is that a principle introduced by the Court in order to find the limit of applicability of the freedom of establishment moves to a reasoning about competition and State aid through the citation. This difference in the citation may be difficult to identify, since the level of semantic similarity of the text is reported as being quite high. In this case the non-semantic score is an effective signal that the citation contains differences. While in itself this is not a reason for dismissing the reference, it may signify the reference deserves further analysis.

Such a comparison is much more evident if we look at the second and third paragraph, where the conceptual distance is bigger, since it loses the reference to the identical tax measure. Furthermore, in this citation chain, the Court also un-anchored the references, since in 2021 any referral to the 1976 judgement disappears.

Furthermore, it must be underlined that the sentence "The same is true in the field of State aid", starts another short chain, in which the same principle may be retrieved in two important cases (ANGED and Gibraltar) pertaining to the field of State aid. However in Gibraltar, which is a leading case, is referred to as an *obiter dictum* coming from the judgement of the Court of first instance. Such precedent is weaker that the ones indicated in the main chain that we analysed. Consequently, it is possible to imagine that Judges chose to un-anchor it, in order to reinforce the case-law support to their statement, which is fundamental for the solution of the case.

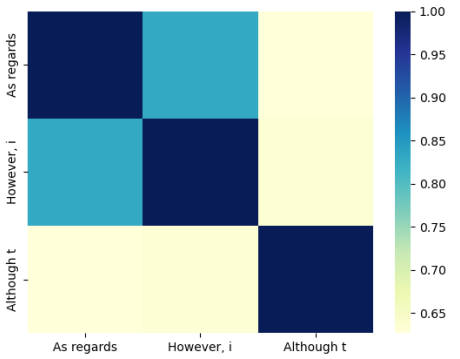


**Figure 4:** Cosine similarity heatmap

A second result of the application of the analysis based on semantic similarity is the possibility to identify the judgement in which the formulation of an interpretative principle reach its "canonic" formulation. From that moment on it really became an "autonomous object". Indeed, when an interpretative principle is "canonized" it became part of the European legal knowledge by itself, detached from the importance of the case law on which it relies.

In figure 5 it is possible to visualize what previously mentioned on *substitution*. The figure maps the degree of semantic similarity between paragraphs that cite one another vertically, from Case ECLI:EU:C:2021:201 to ECLI:EU:C:1990:125[8].

The yellow line in Figure 5 represents, for instance, that at par. 38 of the judgement EU:C:2006:197, the well known Enirisorse case, there is an interpretative turning

[7]The model used is the all-MiniLM-L6-v2, tuned on a 1B sentence pairs dataset from different sources (Wikipedia, Reddit, and Stack Exchange)

[8]The list of paragraphs analysed is: ECLI:EU:C:2021:202, par 33, ECLI:EU:C:2021:201, par 27, ECLI:EU:C:2016:981, par 53, ECLI:EU:C:2015:470, par 24, ECLI:EU:C:2015:235, par 17, ECLI:EU:C:2010:481, par 39, ECLI:EU:C:2006:197, par 38, ECLI:EU:C:2003:415, par 74, ECLI:EU:C:2002:294, par 68, ECLI:EU:C:1990:125, par 25, ECLI:EU:C:1994:325, par 20

point. Such paragraph contains the fundamental statement according to which a measure must be classified as aid only in light of the cumulative fulfilment of all four conditions required by the Treaties. It refers to different precedents, but it recompiles the definition and becomes a canon of interpretation, with small to no semantic differences from that moment on.

While these requirements are mentioned in previous cases, we can verify that the Enirisorse case is where there is a big difference in similarity between citations. The difference has to do with the fact that for the first time there is mention of the requirement of cumulative fulfillment. This in turn is derived from other cited precedents. In this way Paragraph 38 of Enirisorse moves from such precedents and creates a new interpretative principle expressed in a general and abstract manner.

The passage from precedents more strictly connected to the specifics of the cases to this general and abstract formula marks the essential step to the canonization of the interpretative principle. Hence, from the subsequent judgment, EU:C:2010:481 (Deutsche Post AG), there is a significant consolidation of the text, with an average similarity score higher than 0.8.

What that can seems an anomaly, namely the low similarity between Enirisorse and Deutsche Post, is explained by the fact that the latter makes a joint citation of paragraphs 38 and 39 of the first. Which means that the innovative part (the one on cumulative fulfillment) is merged with the descriptive part, the above mentioned chain departing from art. 39 of Enirisorse. So the use of semantic similarity applied to citation network allowed us to find the turning point in the building of a fundamental interpretative principle. Moreover, having identified the critical point through the similarity evaluation, it has been easier to detect which are the steps walked by the Court in reaching its result.

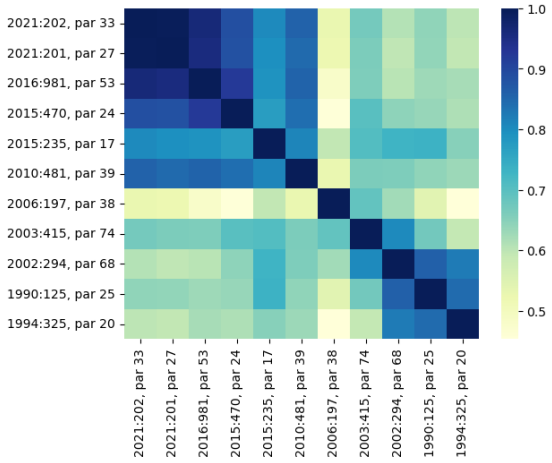


**Figure 5:** Similarity of connected paragraphs

## 5. Future development

While at the moment the automated assessment on large scale is not yet feasible, it is one of the possible future developments, in order to show statistics on the evolution of a particular citation over time. This is not a straightforward task, and a more in depth manual analysis is needed to correctly identify the relevant information.

### 5.1. Semantic similarty

In future work, the similarity analysis could be further developed through the use of domain-specific word embeddings, to see whether their adoption may contribute towards better results. Indeed, some pre-trained models already exist for the legal domain, such as the *Law2Vec* model, which has been trained on legal documents from the EU (including EUR-Lex) and US, although they do not always carry to increases in performance, as described in [9]. Hence, we could develop a semi-automatic comparison between the results of a semantic and non-semantic similarity analysis. To this aim it will be possible to identify immediately potentially critical nodes.

An interesting consideration that can be expanded upon is the difference in performance when using a more semantically-aware measure (by using word- and sentence-level embeddings), taking into account word ambiguity and the different meanings in specific contexts.

We would enhance the analysis of both verbatim and non-verbatim citations by considering the context, in order to verify if there is a substitution. Furthermore we wold try to develop a tool that may highlight inconsitencies in the cited and citing texts.

Another interesting development has to do with the context of the cited cases, and how adding other information could enhance the analysis. In particular identifying whether paragraphs that cite one another, irrespective of the degree of similarity, come from comparable cases. One significant outcome of this could be identifying groups of homogeneous citation (for type of procedure, matter, outcome, etc.) in apparently non-innovative judgement (i.e. a judgment that does not introduce any significant interpretative innovation). With the context information added to the analysis, it could be possible to shed light on trends in legal discourse.

With regard to non verbatim citations, i.e. when the citation changes, refining the similarity analysis may be used to better understand when a principle becomes consolidated, and the case in which it is defined becomes a landmark case, or precedent. In this case a deeper analysis of the different branches, as in Figure 5, could

show when a specific citation becomes a general principle, which is then cited verbatim in newer cases.

Furthermore, the similarity analysis may be used in combination with the community detection method to identify implicit citations, and to compare paragraphs that share only part of the citation chain, or that are indirectly connected, citations that do not directly connect but have a common set of citations.

### 5.2. Community detection

A different metric that can assist in the legal analysis is the community detection, an evaluation and identification of groups (clusters) of nodes and the strength of their grouping.

In this case, it is possible to identify clusters of citations and where they connect or split different branches. In Figure 6 the two metrics (community and centrality) have been shown together, with the colors referring to the clusters and the size of the nodes highlighting their importance within the graph.

The different groupings correspond to the vertical dimension mentioned above, and can be split further. This feature has been used to assist the human analysis and interpretation of the graph, but it could be also linked to some future semantic analysis to better identify *landmark cases* and concepts.

In future developments the community detection could be enhanced with other components, enabling for one a way of filtering the cases in the citation graph.

**Figure 6:** Louvain grouping with Degree centrality.

## 6. Conclusions

When studying and commenting CJEU cases or the assets of the case-law in a specific field, academics are used to reading their impact in the specific context, which means comparing the factual relevant aspects, the state-of-the-art of the discipline both in the literature and in practice. However, this activity is complex and time consuming without the support of automated tools.

In this paper, a methodology for the extraction and analysis of citations and their use with a selection of cases from the European Court of Justice was presented, and applied to a limited domain as a practical use case scenario. In particular, it has been shown that an analysis of the citation network can give useful insights for legal scholars in understanding the vast amount of case law available. In particular, it can be useful *i)* to visualise how the Court uses citations and their different meanings, *ii)* to carry temporal analyses by identifying sequences of citations over time, and *iii)* to capture the relevance of citations by integrating both semantic and network algorithms and metrics.

At present, we worked on a relatively small dataset, though obtaining appreciable results, sometimes related to issues present in the legal studies and debates, as in the case of the interpretative principle about the limit of national sovereignty in the field of tax law.

## References

[1] W. M. Landes, R. A. Posner, Legal precedent: A theoretical and empirical analysis, The Journal of Law and Economics 19 (1976) 249–307.

[2] J. H. Friedman, W. Stuetzle, Projection pursuit regression, Journal of the American statistical Association 76 (1981) 817–823.

[3] W. M. Landes, L. Lessig, M. E. Solimine, Judicial influence: A citation analysis of federal courts of appeals judges, The Journal of Legal Studies 27 (1998) 271–332.

[4] F. B. Cross, T. A. Smith, A. Tomarchio, Determinants of cohesion in the supreme court's network of precedents, San Diego Legal Studies Paper (2006).

[5] J. H. Fowler, T. R. Johnson, J. F. Spriggs, S. Jeon, P. J. Wahlbeck, Network analysis and the law: Measuring the legal importance of precedents at the us supreme court, Political Analysis 15 (2007) 324–346.

[6] M. Derlén, J. Lindholm, Goodbye van gend en loos, hello bosman? using network analysis to measure the importance of individual CJEU judgments, European Law Journal 20 (2014) 667–687. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/eulj.12077. doi:10.1111/eulj.12077.

[7] M. Derlén, J. Lindholm, Is it good law? network

analysis and the cjeu's internal market jurisprudence, Journal of International Economic Law 20 (2017) 257–277.

[8] J. Frankenreiter, Network analysis and the use of precedent in the case law of the cjeu – a reply to derlén and lindholm, German Law Journal 18 (2017) 687–694. doi:10.1017/S2071832200022112.

[9] K. Moodley, P. V. Hernandez Serrano, G. van Dijck, M. Dumontier, Similarity and Relevance of Court Decisions: A Computational Study on CJEU Cases, Legal Knowledge and Information Systems (2019) 63–72. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA190307. doi:10.3233/FAIA190307.

[10] U. Sadl, F. Tarissan, The relevance of the network approach to European (case) law : reflection and evidence, Oxford University Press, 2020. URL: https://cadmus.eui.eu/handle/1814/70596. doi:10.1093/oso/9780198871477.001.0001.

[11] U. Šadl, Old is new: The transformative effect of references to settled case law in the decisions of the european court of justice, Common Market Law Review 58 (2021) 1761–1788. doi:10.54648/cola2021111.

[12] M. Palmirani, R. Brighi, M. Massini, Automated extraction of normative references in legal texts, in: Proceedings of the 9th international conference on Artificial intelligence and law, 2003, pp. 105–106.

[13] E. Leitner, G. Rehm, J. Moreno-Schneider, Fine-grained named entity recognition in legal documents, in: M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, Y. Sure-Vetter (Eds.), Semantic Systems. The Power of AI and Knowledge Graphs, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 272–287. doi:10.1007/978-3-030-33220-4_20.

[14] N. Sakhaee, M. C. Wilson, Information extraction framework to build legislation network, Artificial Intelligence and Law 29 (2021) 35–58. doi:10.1007/s10506-020-09263-3. arXiv:1812.01567 [cs].

[15] P. Zhang, L. Koppaka, Semantics-based legal citation network, in: Proceedings of the 11th international conference on Artificial intelligence and law, ICAIL '07, Association for Computing Machinery, New York, NY, USA, 2007, pp. 123–130. doi:10.1145/1276318.1276342.

[16] A. Sadeghian, L. Sundaram, D. Z. Wang, W. F. Hamilton, K. Branting, C. Pfeifer, Automatic semantic edge labeling over legal citation graphs, Artificial Intelligence and Law 26 (2018) 127–144. doi:10.1007/s10506-018-9217-1.

[17] P. V. Hernandez Serrano, K. Moodley, G. Van Dijck, M. Dumontier, Sleeping Beauties in Case Law, Legal Knowledge and Information Systems (2020) 231–234. URL: https://ebooks.iospress.nl/doi/10.3233/FAIA200871. doi:10.3233/FAIA200871.

[18] A. Louis, G. van Dijck, G. Spanakis, Finding the law: Enhancing statutory article retrieval via graph neural networks, 2023. doi:10.48550/arXiv.2301.12847. arXiv:2301.12847 [cs], type: article.

[19] H. d. Jong, G. v. Dijck, Network analysis in legal history: an example from the Court of Friesland: Remarks on the benefits, Tijdschrift voor Rechtsgeschiedenis / Revue d'histoire du droit / The Legal History Review 90 (2022) 250–262. URL: https://brill.com/view/journals/lega/90/1-2/article-p250_9.xml. doi:10.1163/15718190-20220004.

[20] L. C. Freeman, Centrality in social networks conceptual clarification, Social Networks 1 (1978) 215–239. URL: https://www.sciencedirect.com/science/article/pii/0378873378900217. doi:10.1016/0378-8733(78)90021-7.

[21] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank citation ranking: Bringing order to the web., Technical Report, Stanford InfoLab, 1999.

[22] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, 2019. doi:10.48550/arXiv.1908.10084. arXiv:1908.10084 [cs], type: article.