Ex Uno Plura: Identifying Issues in CJEU Jurisprudence

Philipp Schroeder[1] and Johan Lindholm[2]

[1]Postdoctoral Fellow at Umeå University, Department of Law. Email: philipp.schroeder@umu.se

[2]Professor of Law at Umeå University, Department of Law. Email: johan.lindholm@umu.se

October 20, 2020

Abstract

Research of judges and courts traditionally centres on court decisions, treating each decision as a unit of observation. However, court decisions often address multiple distinct and more or less unrelated issues. Studying judicial behavior on a decision-level therefore loses potentially important details and, more problematically, risks drawing false conclusions from the data. This contribution presents a method to assist researchers with splitting judicial decisions by issues using a supervised machine learning classifier. We illustrate our approach by splitting the CJEU's preliminary rulings concerning the free movement of goods published between 1998 and 2011 into issues. Using this data to replicate a study on the CJEU's strategic references to its own case law, we demonstrate the advantages of choosing issues rather than decisions as units of observation in research of judicial behavior.

1 The problem of multi-issue judgments

While empirical research of courts and judicial behavior has been primarily based on court decisions, the decision is not necessarily the best unit of observation. On the contrary, we argue that court decisions frequently—even typically—address multiple distinct issues and that using issues as a unit of observation is both feasible and advantageous.

Take for example the judgment of the Court of Justice of the European Union (CJEU) in *van Gend en Loos*.[1] The decision is most famous for its constitutional aspects, in particular the Court's characterization of EU law as an autonomous legal order capable of creating individual rights and duties that can be invoked directly on the national level without being implemented by the Member States (see e.g. Chalmers and Barroso 2014). However, in *van Gend en Loos* the Court also addressed and clarified its own jurisdiction,[2] clarified the correct interpretation of the Member States' obligations under EU law with regard to the customs union, and answered whether the Netherlands had failed to comply with those obligations.[3] Lawyers experienced with EU law will say that *van Gend* is not "really" about those latter issues, even though the Court indisputably spent as much ink addressing them as the constitutional issues.[4]

How does the lawyer know what *van Gend en Loos* is "really" about? A lawyer reading the case will have no trouble noticing that the judgment addresses several issues that have little to nothing to do with each other and identifying which paragraphs address which issues. The lawyer will then draw from her knowledge of national, international, and Union law to recognize the exceptional and controversial nature of the Court's declaration of a new legal order, as well as the less significant systemic impact of the Court's findings on the jurisdictional and substantive issues.

Judgment-driven research has struggled to make and include these determinations which are fundamental for understanding the nature and systemic impact of judgments. The last decade has seen a significant rise in the use of citation networks to understand courts and jurisprudence but these studies have generally suffered from a lack of nuanced data (Sadl and Panagis 2015; Winkels et al. 2011). For example, following the approach of studies of other courts (see Fowler et al. 2007; Lupu and Voeten 2012; Winkels et al. 2011), Derlén and Lindholm (2014) studied the CJEU's references to its own previous decisions using network analysis and concluded that the systemic importance of some decisions, such as *Bosman*,[5] has been overlooked. However, this

[1] Judgment of the Court of 5 February 1963. NV Algemene Transport - en Expeditie Onderneming van Gend & Loos v Netherlands Inland Revenue Administration. Reference for a preliminary ruling: Tariefcommissie - Netherlands. Case 26-62.

[2] More specifically, what constitutes a question concerning the interpretation of the Treaties for the purposes of what is now Article 267 of the Treaty on the Functioning of the European Union (TFEU).

[3] The case concerned whether a re-classification of a particular chemical for custom purposes constituted a violation of an article in place at the time that prohibited increased custom duties between Member States.

[4] Speaking from personal experience, this is much less obvious to students that are first learning about EU law.

[5] Judgment of the Court of 15 December 1995. Union royale belge des sociétés de football association ASBL v Jean-Marc Bosman, Royal club liégeois SA v Jean-Marc Bosman and others and Union des associations européennes de football (UEFA) v Jean-Marc Bosman. Reference for a preliminary ruling: Cour d'appel de Liège - Belgium. Case

model approached each decision as a single and indivisible entity and failed to reflect that *Bosman* addresses and is frequently cited by the Court on multiple more or less distinct issues (see Derlén and Lindholm 2016). Thus, while using decisions as the unit of observation correctly captures the central position of *Bosman* in CJEU jurisprudence it fails to acknowledge potentially important structural differences in centrality based on issues.

This lack of nuance in the descriptive data is problematic when it is subsequently analyzed. For example, many judgment-driven studies seek to explore differences (in judicial behavior) based on issues (see e.g. Lupu and Voeten 2012). This may involve the use of manually assigned labels or legislation cited.[6] However, while this may help capture *which* issues a judgment concerns it fails to resolve the problem illustrated with *van Gend en Loos* of determining to *what extent* the case concerns the various topics. Automated textual analysis, such as unsupervised classification using LDA topic modelling (see e.g. Carter et al. 2016; Lindholm 2019; Sadl and Panagis 2015), can help quantify how much of the judgment text deals with a particular topic but this does not necessarily reflect the legal structural significance of the judgment.

Moreover, it does little to improve analyses of differences between issues and the systematic relationship between issues and other factors. For example, Larsson et al. (2017) found that the CJEU embeds its decisions in case law to a greater extent when it faces a more adverse political environment. However, whereas a systematic relationship could be found on a judgment-level, it is not necessarily true that the issues on which it faced political adversity were the same as those on which it referred to more and stronger jurisprudence. A similar problem arises in studies exploring differences between judgments based on areas or topics. Derlén and Lindholm (2015) previously explored the claim in the literature that the role of CJEU jurisprudence differs significantly depending on the subject matter. The study centred on the correlation between the case law cited in a decision with the Court's own subject matter classification, a problem being that the average decision involves two to three such subject matters.

In this article, we demonstrate how text classification techniques commonly used in the field of natural language processing can facilitate splitting court judgments into issues and avoid the problems associated with decision-based research of judges and courts. The article proceeds as

C-415/93.

[6] One commonly used type being court-provided subject matters.

follows. The following section conceptualizes legal issues as units of observation in research of judicial behaviour and discusses uses of splitting court judgments into issues. The third section introduces an approach relying on supervised machine learning classification to facilitate the issue-splitting of larger samples of court decisions. We illustrate the approach by splitting the CJEU's judgments on preliminary references concerning the free movement of goods published between 1998 and 2011 into their issues. In the fourth section we make use of this data to replicate analyses conducted by Larsson et al. (2017) on the CJEU's strategic references to its own case law, but rely on issues rather than judgments as our units of observation. Following a brief discussion of our results, the final section offers concluding remarks.

2 Issues as units of observation and a Goldilock layer

Working closely with the text on a detailed level improves our ability to understand how courts behave and why (Sadl and Panagis 2015). However, treating each sentence or paragraph of a judgment as a unique, unconnected entity is also sub-optimal as it fails to acknowledge that sentences and paragraphs deal with and are connected through a limited number of legal questions. The solution, we argue, is to split the judgment text into blocks that deal with a particular legal question. These blocks, which we refer to as *issues*, are essentially a type of clustering data applied to judgment sentences or paragraphs.[7] The issue layer serves as a connecting middle layer between the judgment level and the paragraph level, clustering paragraphs addressing the same legal question (see Figure 1).

This has several uses. First, we can classify what topic the issue concerns by analyzing the associated text and/or the references, e.g. using unsupervised approaches such as topic modelling or network-based clustering on the basis of reference. Thus, we distinguish between the clustering problem of what text deals with the same legal question (*issues*) and the classification problem of determining the nature of that question (*topics*). While there is no objectively correct number of topics, in order for topics to be useful analytical tools there should obviously be significantly fewer topics than issues.

[7] In the case of our data, the jurisprudence of the CJEU, paragraphs are appropriate as the smallest units of text as they are relatively short, internally homogeneous in terms of subject and voice, easily identifiable through a unique court-assigned identifier, and used by the Court when referring to its own jurisprudence. A different division may be more appropriate when studying another court.

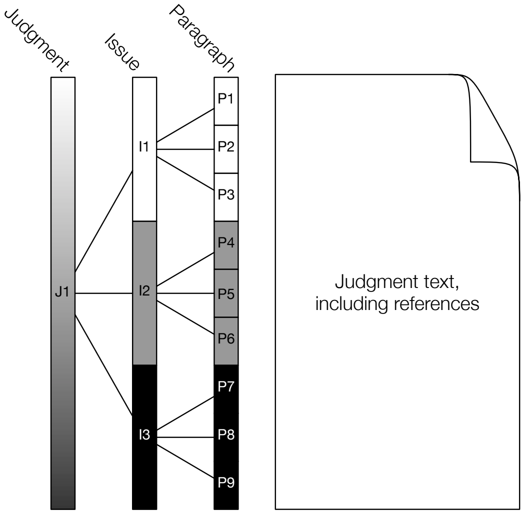

Figure 1: A simple example of a judgment (J1) containing nine paragraphs of text (P1—P9), including references to sources, dealing with three legal questions and clustered into three issues (I1–I3). Whereas we can determine that the judgment includes discussion about three questions, the issue layer allows us to determine to what extent and which text relates to which question.

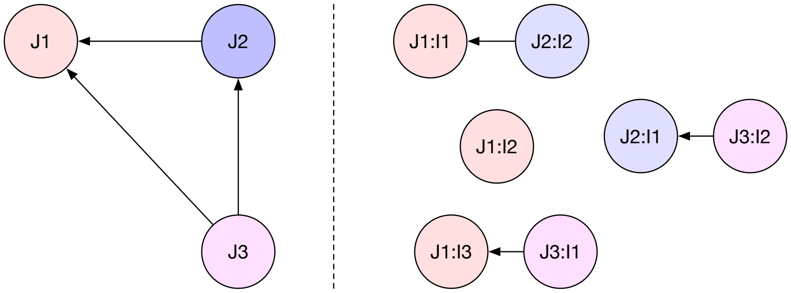

Figure 2: On the left, an example of a judgment-to-judgment network containing three judgments where the middle one (J2) contains a reference to the oldest (J1) and the newest (J3) contains references to both. On the right, an example of a issue-to-issue network based on the same judgments and references but split by issue.

Second, this allows case-law references to be analyzed on an issue level and for the construction of issue-to-issue citation networks. Such a network would more accurately capture relevant references between judgments that allows for more accurate representation and analysis of network structure and centrality (see Figure 2). It would also make it possible to assess how central a judgment, such as *van Gend en Loos* or *Bosman*, is in the context of a particular topic.[8]

[8] While network analysis allows us to capture the relative centrality of a judgment within a cluster, this overcomes the problem that the entire judgment will belong to a single cluster.

3 A machine learning assisted approach to issue-splitting

Lawyers familiar with the jurisprudence of a particular court are undoubtedly capable of identifying and splitting a decision by the different issues addressed by a court. However, this task is time consuming and practically unfeasible in many projects, in particular those involving a large number of decisions and if one seeks to keep the dataset updated. To mitigate the time and resources otherwise required to identify separate issues over a large number of court decisions, we suggest a supervised machine learning assisted approach to issue-splitting.

3.1 Supervised classification using recurring language patterns

While a host of words contained in the paragraphs that make up the text of a court decision are likely to be associated with the subject matter of the case, we often find that courts consistently use certain words and expressions that allow us to identify different classes of paragraphs without having to understand the meaning of (or in fact, even read) these paragraphs.

To illustrate, consider the following excerpts from a typical preliminary reference case heard by the CJEU. In *Agostinho da Silva Martins v Dekra Claims Services Portugal SA*,[9] a national court required clarification from the CJEU regarding a possible conflict between a supranational regulation and national provisions covering damage compensation claims. The CJEU identified two separate issues from the questions referred by the national court, introducing the first issue in the following paragraph:

> By its first question, the referring court is asking, in essence, whether Article 16 of the Rome II Regulation must be interpreted as meaning that a national provision, such as that at issue in the main proceedings, which provides that the limitation period for actions seeking compensation for damage resulting from an accident is three years, may be considered to be an overriding mandatory provision, within the meaning of that article.

After addressing the national court's question and stating an answer in subsequent paragraphs, the CJEU then introduced the second issue as follows:

[9] Judgment of the Court (Sixth Chamber) of 31 January 2019. Agostinho da Silva Martins v Dekra Claims Services Portugal SA. Request for a preliminary ruling from the Tribunal da Relação de Lisboa. Case C-149/18.

> By its second and third questions, which it is appropriate to examine together, the referring court is asking, in essence, whether Article 27 of the Rome II Regulation must be interpreted as meaning that Article 28 of Directive 2009/103, as transposed into national law, constitutes a provision of EU law which lays down a conflict-of-law rule relating to non-contractual obligations, within the meaning of Article 27 of that regulation.

While these paragraphs introduce separate issues, both contain expressions such as “asking, in essence”, “by its [...] question” and “must be interpreted as meaning”. As illustrated further below, we found that these similarities across paragraphs serving the same function in the CJEU’s jurisprudence (here, introducing the CJEU’s answer to a question referred by a national court) are not merely incidental but are by and large systematic. Where courts consistently use certain patterns of language in paragraphs serving the same function within their decisions, we can employ supervised machine learning to assist with the issue-splitting of court decisions on a larger scale.

Our approach to issue-splitting comprises four steps. First, we select a sample from a population of court decisions and parse the text of these decisions into paragraphs. In the second step, a coder reads these paragraphs and assigns one out of a predefined set of classes to each paragraph (e.g. a class indicating the paragraph in question marks the beginning of a court’s discussion of an issue, or a class identifying the court’s concluding paragraph on an issue). Third, we split the hand-coded paragraphs into a training and validation set, using the former to train a supervised machine learning classifier suitable for text classification and evaluate the classifier’s performance on the validation set. Finally, the best-performing classifier is then used to predict paragraph classes for court decisions that have not been used to train the classifier, i.e. the test set. Classified paragraphs allow for court decisions in the test set to be split into their issues. To illustrate, recall Figure 1. A text block comprising an issue simply begins at the paragraph marking the start of the court’s discussion of an issue, ends at the next paragraph marking the conclusion of the court’s discussion of that issue, and comprises all paragraphs between these two.

In the following, we discuss the performance of a supervised machine learning classifier for paragraphs from a sample of the CJEU’s preliminary rulings published between 1998 and 2011. A discussion of the training, selection and validation of the supervised machine learning classifier as well as an illustration of improvements in classifications as the volume of training data increases is provided in section A of the appendix.

3.2 Classifying paragraphs in CJEU jurisprudence

For the purpose of our illustration, we are interested in issue-splitting the CJEU's preliminary rulings on questions referred by national courts concerning the free movement of goods that the Court delivered between 1998 and 2011. The CJEU published 205 such judgments within this time frame, comprising our test set. Our training data for the classifier comprises roughly 9,000 paragraphs from a sample of more than 400 preliminary reference judgments the CJEU issued between 2006 and 2019 (for details, see section A.1 in the appendix).

In order to split the CJEU's preliminary rulings into issues, we distinguish between four different paragraph classes: (1) paragraphs introducing the CJEU's response to a national court's referred question (`question_start`), (2) paragraphs stating the CJEU's concluding response to the national court's question (`question_stop`),[10] (3) paragraphs stating that a national court question does not require an answer from the CJEU (`question_noanswer`),[11] and (4) a residual category for all remaining paragraphs in the judgment (`arguments`). A machine learning classifier can be expected to perform reasonably well in distinguishing between paragraph classes that are characterized by distinct language patterns. Table 1 lists the most frequent stemmed 3-grams found in the training data for each class (i.e. allowing us to capture multi-word language patterns), showing the promise of using supervised machine learning to classify paragraphs in the CJEU's preliminary rulings. The listed most frequent 3-grams for the classes `question_start`, `question_stop` and `question_noanswer` have a plausible connection to their respective paragraph classes, while 3-grams for the class `arguments` do not appear to follow a particular pattern, precisely what we would expect from a residual category.

[10] A typical example of a paragraph of the class `question_stop` reads "In view of the foregoing considerations, the reply to the first question is that the separation, crushing and purification of silicon metal blocks and the subsequent sieving, sorting and packaging of the silicon grains resulting from the crushing, as carried out in the main proceedings, do not constitute origin-conferring processing or working, for the purposes of Article 24 of the Customs Code", see Judgment of the Court (Third Chamber) of 11 February 2010. Hoesch Metals and Alloys GmbH v Hauptzollamt Aachen. Reference for a preliminary ruling: Finanzgericht Düsseldorf - Germany. Case C-373/08.

[11] A typical example of a paragraph of the class `question_noanswer` reads "Given the answer to the third question, there is no need to answer the fourth question", see Judgment of the Court (Second Chamber) of 19 March 2009. Mitsui & Co. Deutschland GmbH v Hauptzollamt Düsseldorf. Reference for a preliminary ruling: Finanzgericht Düsseldorf - Germany. Case C-256/07.

Table 1: Common features for paragraph classes

Paragraph class	Most frequent features (3-grams)
question_start	the_refer_court, be_interpret_as, must_be_interpret, refer_court_ask, in_essenc_whether
question_stop	the_answer_to, answer_to_the, must_be_interpret, be_interpret_as, to_the_question
question_noanswer	there_is_no, is_no_need, no_need_to, to_the_first, question_there_is
arguments	eu:c_paragraph_and, the_main_proceed, in_the_main, to_that_effect, see_to_that

Note: Paragraph classes per issue type. The column 'Most frequent features (3-grams)' shows the five most frequent 3-grams per paragraph class for the training set (N8,816).

We opt for a neural network to solve our classification problem, a deep-learning algorithm implemented through the **keras** and **tensorflow** packages for the programming language R. Details on the specification of the neural network and its performance on the training data are provided in section A.3 of the appendix. Prior to training the classifier, we removed a list of common stop words, stemmed the remaining words and tokenized them into 3-grams.

We validated (and where necessary, corrected) paragraph classifications by hand, allowing us to report on the classifier's performance on the test set.[12] Table 2 shows that the neural network's classification performance varies across the different paragraph classes. The best classification results are achieved for the residual class arguments, which does not surprise given that it is by far the most frequent paragraph class in both the training and test sets (see section A.1 in the appendix). At the lower end of its classification performance is the class question_stop. Here, about 79% of the predictions made by the model are in fact correct, while it captured 65% of paragraphs which should have been classified as question_stop. The classifier's precision is far better for paragraph classes question_start and question_noanswer, correctly classifying more than 90% of paragraphs for these classes, although arguably at the expense of its recall performance.

Classification problems such as the one discussed here typically face a trade-off between precision and recall (i.e. higher precision can be achieved at the expense of recall, and vice versa). In addition,

[12] Precision refers to the proportion of predictions (per class) which were predicted correctly by the model, i.e. true positives/(true positives + false positives). Recall refers to the proportion of actual instances of a class the model predicted correctly, i.e. true positives/(true positives + false negatives).

Table 2: Classification performance for paragraph classes in test set (N11,077)

Paragraph class	*Precision*	*Recall*	*Freq.*	*Prop.*
question_start	0.91	0.57	390	0.04
question_stop	0.79	0.65	390	0.04
question_noanswer	0.90	0.71	61	0.01
arguments	0.96	0.99	10,236	0.92

Note: Precision and recall rates for paragraph class predictions of a neural network with two hidden layers comprising 480 nodes each. The neural network was trained for two epochs on training data comprising 8,816 paragraphs.

the performance pattern of high precision and relatively low recall rates observed here is likely due to language patterns specific for paragraph classes that had not been picked up in the training data, as well as the CJEU deviating from its usual use of language patterns in some judgments in the test set.[13] Nonetheless, the classification of paragraphs using supervised machine learning can greatly facilitate a coder's task of splitting court decisions by their issues. For our purposes, we programmed an online application that allowed a coder to view judgment texts segmented by paragraphs along with the paragraph class predicted by the deep-learning classifier. Rather than having to input a class for each paragraph, the coder's task was therefore reduced to spotting and correcting mistaken classifications.[14]

Following the validation of the classifier's predictions in the test set, we find that the CJEU addressed 390 separate legal issues in its 205 preliminary rulings on national court questions concerning the free movement of goods published between 1998 and 2011. Further, we find that the CJEU concluded in 61 instances that questions referred by a national court required no answer from the Court. In the following section, we make use of this data to replicate a previous study by Larsson et al. (2017) on the CJEU's strategic references to its own case law, and demonstrate the advantages of opting for issues rather than decisions as units of observation in studies of courts

[13] For instance, in its judgment in the Case C-55/06 the CJEU introduced its answer to a national court question using the following language pattern: "By Question 3(f) the Court of Justice is essentially invited to consider [...]" (see Judgment of the Court [Fourth Chamber] of 24 April 2008. Arcor AG & Co. KG v Bundesrepublik Deutschland. Request for a preliminary ruling from Verwaltungsgericht Köln - Germany. Case C-55/06). This pattern deviates from the common features listed in Table 1, complicating accurate classification, particularly where the volume of available training data is relatively low.

[14] An additional benefit of validating classifications in the test set is that validated paragraphs can be used as additional training data to improve a classifier's accuracy for future applications, a particularly useful feature for researchers interested in keeping their datasets updated as courts continually publish new judgments (see our illustration in section A.4 of the appendix).

and judicial behavior.

4 Application: The CJEU's strategic references to case law

A prominent strand of research in the field of judicial politics contends that courts' behavior is shaped by political constraints. These constraints come in the form of court-curbing in response to judicial decisions thwarting the will of political majorities (see Clark 2010; Whittington 2003; Voeten 2020), authorities' non-compliance and legislative override of unfavorable court rulings (Carrubba and Zorn 2010; Vanberg 2005; Larsson and Naurin 2016), as well as political principals' selective appointment of judges (Voeten 2007).

Existing literature suggests that courts strategically show deference to their political principals when threats of court-curbing or legislative override are credible (see Clark 2009; Carrubba et al. 2008; Larsson and Naurin 2016). However, several studies argue that courts have a variety of tools at their disposal to mitigate the risks of backlash and non-compliance. For instance, Owens et al. (2013) argue that justices at the U.S. Supreme Court obfuscate the language used in their opinions to evade Senate scrutiny. Further, Staton (2006) shows that the Mexican Supreme Court selectively promotes its merits decisions through press releases when authorities' non-compliance is likely.

Adding to the growing volume of scholarship on the CJEU's strategic decision-making, Larsson et al. (2017) discuss another tool available to judges navigating adverse political environments. They argue that the CJEU uses legal justifications as a legitimation strategy when its rulings run counter to the expressed interests of EU Member States. Key findings from their study suggest "that the Court argues more carefully, by means of reference to precedent, when it takes decisions that conflict with the positions of EU governments" (Larsson et al. 2017, 881).

Their empirical analysis draws on two separate datasets. Derlén and Lindholm (2014) identified the CJEU's references to its own case law for all 9,125 judgments the Court had issued between 1954 and May 2011, and capture citation patterns between these judgments using network analysis. This data, which allows for the measurement of variation in the CJEU's reference to precedent across judgments, is complemented by information on the CJEU and EU Member States' positions on the questions the Court discussed in its preliminary rulings issued between 1998 and 2011 (Naurin et al. 2015). While the units of observation for the CJEU's references to precedent in the data provided by Derlén and Lindholm (2014) are judgments, actors' positions are expressed for the specific questions national courts had referred to the CJEU, with a single CJEU-judgment typically dealing with multiple national court questions. Larsson et al. (2017) solve this discrepancy in units of observations by aggregating data on actors' positions to the judgment-level.

Given that the CJEU often considers several more or less related questions in its judgments (see Derlén and Lindholm 2016), this strategy invites criticism that the inferences drawn from the judgment-level data may be misleading. For example, hoping to steer the CJEU's jurisprudence in a favorable direction, Member States may be more likely to express their positions on national court questions where existing case law is thin, but choose to save themselves the trouble of submitting observations to the Court when a well-established body of relevant case law already exists. When relying on aggregated judgment-level data, researchers would struggle to detect such patterns in their analyses. Following the splitting of the CJEU's preliminary rulings into issues, we are in a position to identify the CJEU's references to case law at the issue-level and determine whether such criticism is warranted.

4.1 Operationalizations of outcome and explanatory variables

We are interested in explaining variation in an outcome variable that captures the CJEU's decisions to embed rulings in existing case law. In a first step, we identify the CJEU's citations of its previous case law in the texts of the preliminary rulings using regular expressions.[15] Following our issue-splitting approach providing us with the text blocks for each issue, we can identify citations both at the judgment-level and the issue-level.

Like Larsson et al. (2017), we then construct a variable *Outdegree*, which counts the number of outward citations for a particular unit of observation (i.e. a judgment or an issue). We consider alternative measures to capture the extent to which the CJEU's rulings are embedded in its existing case law, such as the *Hub Score*, a more sophisticated network centrality measure weighing outward citations based on the precedential authority of the cited units (see Easley and Kleinberg 2010; Lupu and Voeten 2012), as our outcome variable in analyses presented in the appendix.

To evaluate whether the CJEU is more likely to embed its decision in existing case law when

[15] Citations of case law in the CJEU's jurisprudence follow a consistent pattern, referring to the cited case name (e.g. *Costa v ENEL*), often followed by a reference to a particular paragraph in that judgment. Using regular expressions we are able to capture both which judgment as well as which paragraph is cited.

the Court's rulings go against the interest of (coalitions of) EU Member States, Larsson et al. (2017) construct a variable *MS Conflict*, comprising three categories: (1) *In conflict*, indicating that the CJEU's ruling touched upon Member States' autonomy while Member States' net position on the ruling was in conflict with the CJEU's position on the issue; (2) *In favor*, indicating that Member States' net position aligned with the CJEU's position on a ruling concerning national autonomy; and (3) *Ambivalent* indicating that either the CJEU or Member States' position did not touch upon the latter's autonomy.[16] The same approach was used to measure whether or not the CJEU's positions conflicted with positions of the Advocate General and the European Commission, captured by the variables *AG Conflict* and *Commission Conflict*, respectively. We reconstruct the variables *MS Conflict*, *AG Conflict* and *Commission Conflict* for our subset of preliminary rulings, both at the judgment-level and issue-level.[17]

Figure 3 plots the distributions for the outcome variable *Outdegree* and the explanatory variable *MS Conflict* at the judgment-level and the issue-level. The top panels in Figure 3 show that aggregating data on the CJEU's references to existing case law at the judgment-level masks the fact that the CJEU typically references only a handful of its previous decisions (if any) when discussing the actual issues at stake in the case. Further, the bottom panels in Figure 3 indicate that splitting the CJEU's preliminary rulings into issues reveals that for most of the issues considered, the CJEU and EU Member States' positions on further restrictions of national autonomy in favor of legal integration were either unclear, or the issue simply did not concern Member States' autonomy.

4.2 Estimation

Our judgment-level data comprises a total of 205 judgments, while our issue-level data comprises the 405 distinct issues discussed by the CJEU within these judgments. The issue-level data has a hierarchical structure, as issues are nested in judgments. However, not every judgment actually

[16] To arrive at Member States' net position on an issue, positions were weighted by the voting power of the member states in the Council, for full details regarding the construction of the variable, see Larsson et al. (2017, 893).

[17] Note that as mentioned above, Member State positions on the issues discussed by the CJEU were originally coded for each question a national court had referred to the CJEU. The CJEU occasionally answers several national court questions collectively within a text block which we would consider an issue in line with our discussion in Section 2. Hence, the original units of observation for the data on actors' positions do not always perfectly match our concept of issues. While we identified a total of 451 issues in our subset of preliminary rulings, data provided by Naurin et al. (2015) indicates that the CJEU dealt with 486 national court question in this subset of CJEU rulings. We matched the positions coded by Naurin et al. (2015) to the issues we identified, matching multiple positions to a single issue and aggregating to the issue-level where necessary.

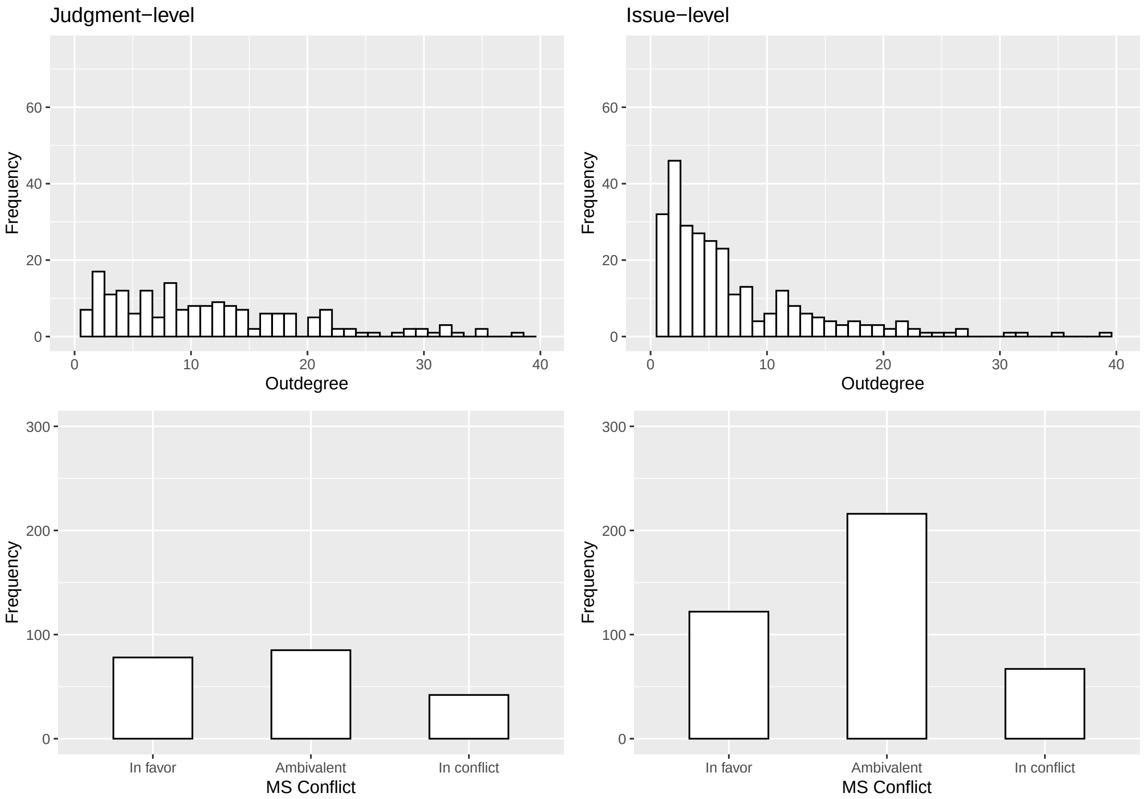

Figure 3: Distributions for the outcome variable *Outdegree* and the main explanatory variable *MS Conflict* at the judgment-level (N205) and issue-level (N405).

comprises multiple issues. Out of the 205 judgments, 105 contain only one issue (hence there is no difference between the judgment- and issue-level data), while we find two or more issues discussed in the remaining judgments (with a maximum of eleven issues). Adding to this complexity, Larsson et al. (2017) include several control variables in their analysis, which are by design measured at the judgment-level. These include the number of observations Member States had submitted (*Sum of Observations*), whether or not the judgment involves the interpretation of Union primary law (*Treaty*), a variable *Complexity* which captures the number of questions submitted by national courts and the number of legal acts the judgment concerned, the number of judges that decided the case (*Chamber Size*), whether or not a court from a common law country submitted the reference (*Common Law Origin*), and fixed-effect controls for the *Year* the judgment was published.

Faced with this type of data, researchers may choose between estimating issue-level regression models, ignoring variation across judgments beyond the judgment-level controls included in the model, and estimating judgment-level regressions after aggregating values for the issue-level predictors and hence removing the ability to let issue-level predictors account for issue-level outcomes. To avoid the downsides of either of these approaches, we suggest to incorporate the hierarchical structure of the data in our models and estimate a multi-level regression that can easily handle predictors at both issue- and judgment-level while accounting for judgment-level variation by allowing intercepts to vary across judgments.

Given our explanatory variable is a discrete count variable, we estimate a negative binomial multi-level regression model. In light of the relatively large number of model parameters that need to be estimated (e.g. 205 random effects, one for each judgment), we can reasonably expect convergence issues when relying on maximum likelihood estimation, particularly given the relatively small sample size of our data. We follow advice by Gelman and Hill (2007) and opt for a Bayesian approach to estimate the model's parameters, specifying uninformative priors for the parameters and running four chains with 10,000 sampling iterations to estimate the parameters' posterior distributions. All estimations are implemented through the **rstanarm** package for the programming language R.

4.3 Results

In the following, we compare evidence from various regression analyses to evaluate empirical support for the argument that the CJEU is more likely to embed its rulings in existing case law when it faces an adverse political environment. Specifically, we estimate three models: (1) a judgment-level regression, aggregating values for the explanatory variables measured at the issue-level for each judgment, (2) a pooled issue-level regression, ignoring variation across judgments, and (3) a partially pooled multi-level regression with varying intercepts for each judgment.

Figure 4 plots the regression coefficients' posterior means along with their 95% highest probability density (HPD) intervals for the three models. Note that the reference categories for the three variables *MS Conflict*, *AG Conflict* and *Commission Conflict* are observations indicating no conflict between the CJEU's position and the positions of Member States, the Advocate General and the Commission, respectively.

We can quickly spot two patterns in Figure 4. First, coefficient estimates for the variables *MS Conflict*, *AG Conflict* and *Commission Conflict* differ—in some instances, substantially—between the judgment-level regression on the one hand, and the issue-level regressions on the other. This

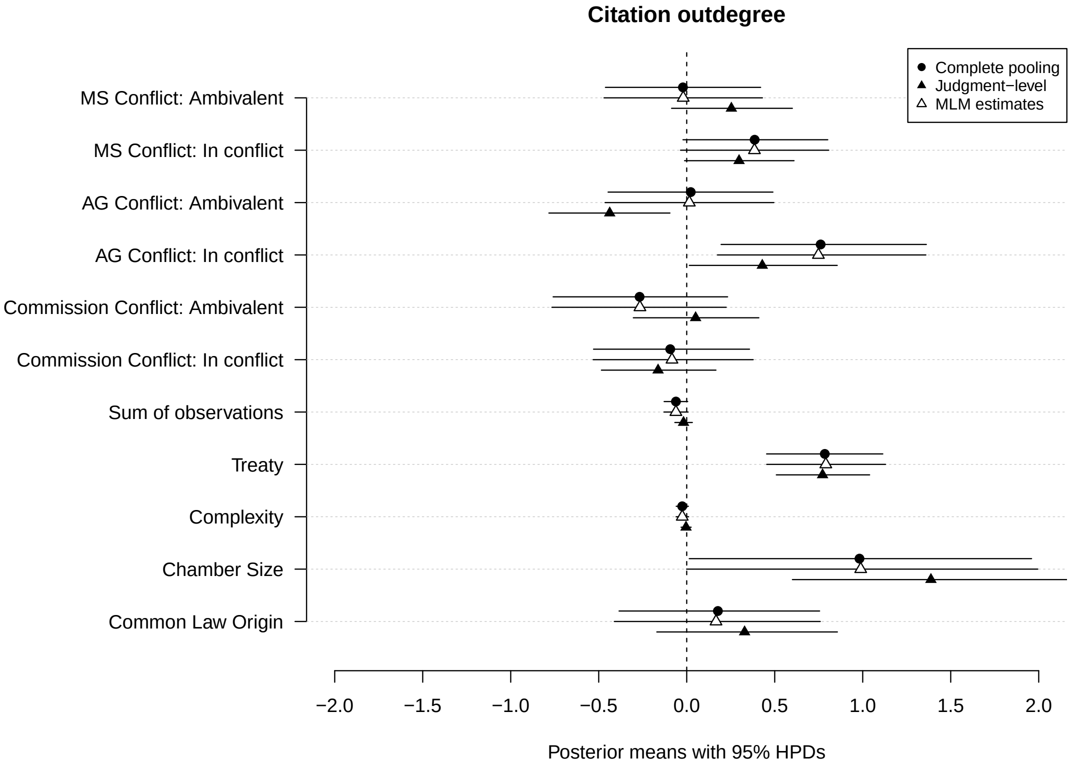

Figure 4: Posterior means with 95% HPD intervals of regression coefficients, displayed for the judgment-level (N205), issue-level and multi-level analyses (N405). All regression analyses include year fixed-effects (not shown here).

is most clearly seen for the category *Ambivalent* of the variable *AG Conflict.* The issue-level regressions indicate no systematic differences in the CJEU's references to existing case law between issues where the Advocate General and the Court held the same position and issues which did not touch upon Member States' autonomy. In contrast, when estimating the regression using judgment-level data with actors' positions on issues aggregated for each judgment, the coefficient estimate suggests that the CJEU is less likely to embed its decision in existing case law when the Court or Advocate General's positions did not concern Member States' autonomy, than when the CJEU shared the Advocate General's position. Our issue-level analyses allow us to conclude that this result (as well as similar patterns for the variables *MS Conflict* and *Commission Conflict*) are artefacts of aggregating actors' positions to the judgment-level and the nuance in information that is lost in the process.

Second, coefficient estimates for the pooled issue-level analysis and the multi-level model only marginally differ between the two approaches, suggesting that accounting for judgment-level vari-

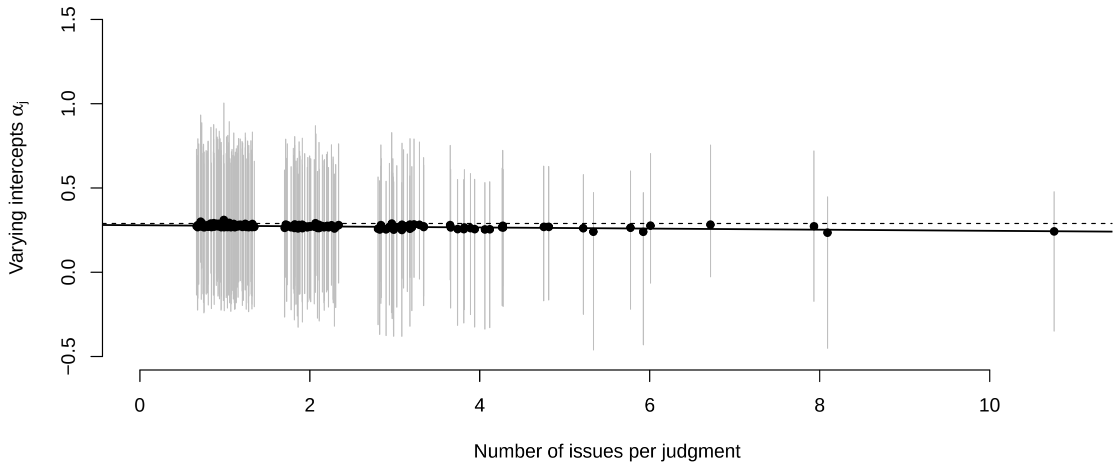

Figure 5: Partial pooling estimates with 5th and 95th percentiles of their posterior distributions for judgment intercepts, plotted against the number of issues discussed in judgments. The dashed line indicates the intercept estimate for the pooled issue-level regression.

ation beyond the control variables included in both models has only minor implications (although there is no harm in estimating the multi-level regression). Corroborating this conclusion, Figure 5 shows that intercept estimates tend to fall close to the pooled estimates for judgments with fewer issues, although we can see that intercepts tend to be marginally smaller for judgments comprising higher numbers of issues.

Beyond these patterns, the substantially most interesting finding of our analyses is that coefficient estimates for the category *In conflict* of the variable *MS Conflict* is positive and distinguishable from zero across all three models. In other words, evidence first uncovered by Larsson et al. (2017) from the judgment-level analysis suggesting that the CJEU is more likely to embed its decisions in existing case law when it takes positions conflicting with the views of Member States remains robust for our issue-level analyses. Based on coefficient estimates from our multi-level issue-level regression, we find that the CJEU on references on average an additional 2.51 judgments from its existing case law when the implications of the Court's decision run counter to the expressed interests of Member States, relative to scenarios in which the Court and Member States share the same position on the issue (see Figure 6).

We find similar results for the category *In conflict* of the variable *AG Conflict*, with coefficient

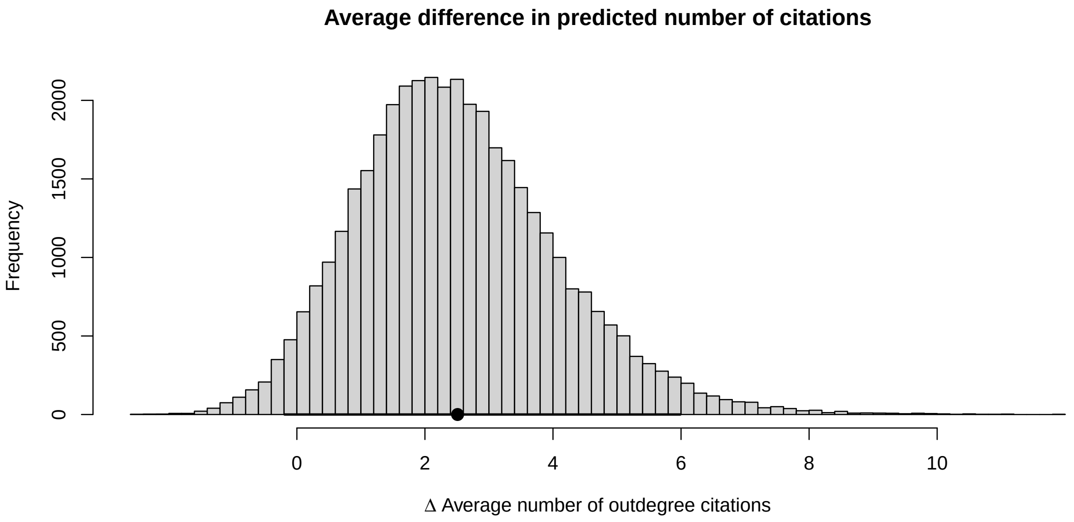

Figure 6: Posterior distribution of the average difference in the predicted number of unique outdegree citations for *MS Conflict = In conflict − MS Conflict = In favor*. Point estimate is shown with the 2.5th and 97.5th percentiles of its posterior distribution.

estimates being positive and clearly distinguishable from zero for all three models, with a much stronger effect once we consider the CJEU's outdegree citations at the issue-level. Substantively, this evidence suggests that the CJEU tends to make a greater effort to embed rulings in existing case law via references to precedent when its decisions run counter to the expressed positions of the Advocate General, the Court's principle advisor. We find no such effect for conflicts between the CJEU's positions and the interests of the European Commission.

To summarize, our replication analysis relying on units of observation at the issue-level rather than the judgment-level reinforces evidence in support of Larsson et al.'s main hypothesis. The CJEU appears more likely to embed its rulings in existing case law when its position on questions concerning national autonomy conflicts with the interests of Member States and the position expressed by the Advocate General, than when the Court's position is shared by those actors. This evidence is consistent with expectations that the CJEU uses references to existing precedent as a legitimation strategy when navigating adverse political environments. Our analysis shows that this finding is robust when considering a significantly smaller sample of judgments and issues than Larsson et al.'s original analysis, and preliminary rulings that deal only with questions concerning intra-EU free movement of goods. However, beyond this pattern we were able to show that aggregating data that is available at the issue-level to the judgment-level and the resulting loss of information are likely to lead researchers to draw misleading inferences from their analyses.

5 Conclusion

Legal scholars and scholars of judicial politics have urged students of judicial behavior to centre their attention on the text of courts' jurisprudence (see for example Lax 2011). Tiller and Cross (2006, 523) argue that "[t]he language of the opinion at least purports to establish the rules to govern future cases, but political science researchers have generally disregarded the significance of this language". Empirical analyses of judicial behavior, particularly those involving larger datasets, have tended to reduce courts' decisions to dichotomous outcomes, overlooking that "decisions are often most important because of the qualitative changes in law that they effect, rather than because of the decision they provide on the case facing the Court" (Clark and Lauderdale 2010, 871).

However, once scholars shift their attention away from outcomes and towards the language of court decisions, they are faced with large volumes of text from judgments that commonly address multiple issues. In addition to resolving disputes over substantive legal issues, courts at the helm of their judicial hierarchies often consider questions of admissibility of the questions before them and their jurisdiction over the resolution, and in some instances—such as for the CJEU—are explicitly concerned with a variety of more or less related substantive legal issues within individual judgments.

In this contribution, we introduced an approach that reduces such complexity of judgments and allows researchers to structure the text of judgments into clusters of paragraphs that deal with distinct, internally consistent issues. We outlined a machine learning assisted approach that facilitates the splitting of judgments into issues, using text classification techniques and recurring language patterns in judgment texts. Although our approach does not eliminate the need for manual coding, it reduces the time and effort coders would otherwise need to identify distinct issues in judgments and appears particularly promising for scholars interested in keeping their datasets updated as courts continually publish new jurisprudence, given that classification results are likely to improve as more and more training data becomes available.

A key benefit of our approach of classifying paragraphs to provide structure to judgments is that researchers end up with the actual text for each issue that is discussed in a judgment. This opens up a variety of new opportunities for empirical research. In our application discussed above, we were able to identify the CJEU's references to precedent from texts specific to issues the Court considered in its judgments, and field more nuanced data to evaluate existing expectations about the CJEU's strategic behavior. But this is by far not the only application we may think of. Rather than having to rely on full judgment texts, which often include more information than we care for in a particular research project, scholars may for instance construct measures connected to relevant aspects of judicial behavior based on word counts, lexical diversity or sentiment analyses specific to each substantive issue a court considered in its judgment.

Our experience of splitting a subset of the CJEU's preliminary rulings into issues suggests that supervised machine learning classifiers allow us to provide structure to complex judicial decisions without having to read every single word within them, although we are conscious that the context of preliminary references proceedings, with national courts submitting distinct legal questions for the CJEU to resolve, appears particularly well-suited to our approach. Nonetheless, we are confident that our suggested approach can be used (or modified) to identify similar structures and coherent units of text in the jurisprudence of other courts as well. Whenever courts regularly use certain patterns of language in their judgments, researchers can employ machine learning classifiers trained to identify such patterns to provide structure to large volumes of previously unstructured text. Our approach then helps to reduce the complexity of courts' jurisprudence, which would otherwise present obstacles to text-driven research of judicial behavior.

References

Carrubba, Clifford J. , Matthew Gabel, and Charles Hankla (2008). Judicial Behavior Under Political Constraints: Evidence from the European Court of Justice. *American Political Science Review 102*(04), 435–452.

Carrubba, Clifford J. and Christopher Zorn (2010). Executive Discretion, Judicial Decision Making, and Separation of Powers in the United States. *The Journal of Politics 72*(03), 812–824.

Carter, David J. , James Brown, and Adel Rahmani (2016). Reading the High Court at a Distance: Topic Modelling the Legal Subject Matter And Judicial Activity of the High Court of Australia, 1903–2015. *University of New South Wales Law Journal 39*, 1300–1354.

Chalmers, Damian and Luis Barroso (2014, 04). What Van Gend en Loos stands for. *International Journal of Constitutional Law 12*(1), 105–134.

Clark, Tom S. (2009). The Separation of Powers, Court-Curbing and Judicial Legitimacy. *American Journal of Political Science 53*(4), 971–989.

Clark, Tom S. (2010). *The Limits of Judicial Independence*. Cambridge: Cambridge University Press.

Clark, Tom S. and Benjamin E. Lauderdale (2010). Locating Supreme Court Opinions in Doctrine Space. *American Journal of Political Science 54*(4), 871–890.

Derlén, Mattias and Johan Lindholm (2014). Goodbye van Gend en Loos, Hello Bosman? Using network analysis to measure the importance of individual CJEU judgments. *European Law Journal 20*(5), 667–687.

Derlén, Mattias and Johan Lindholm (2016). Bosman: A Legacy Beyond Sports. In A. Duval and B. Van Rompuy (Eds.), *The Legacy of Bosman*, pp. 31–49. The Hague: T.M.C. Asser Press.

Derlén, Mattias and Johan Lindholm (2015). Characteristics of Precedent: The Case Law of the European Court of Justice in Three Dimensions. *German Law Journal 16*(5), 1073–1098.

Derlén, Mattias and Johan Lindholm (2017). Is it good law? network analysis and the cjeu's internal market jurisprudence. *Journal of International Economic Law 20*, 257—-277.

Easley, David and Jon Kleinberg (2010). *Networks, crowds, and markets.* Cambridge: Cambridge University Press.

Fowler, James H. , Timothy R. Johnson, James F. Spriggs, Sangick Jeon, and Paul J. Wahlbeck (2007). Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. *Political Analysis 15*(3), 324–346.

Gelman, Andrew and Jennifer Hill (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Kleinberg, Jon M. (1999). Hubs, authorities, and communities. *ACM Computing Surveys 31*, 604—-632.

Larsson, Olof and Daniel Naurin (2016). Judicial Independence and Political Uncertainty: How the Risk of Override Affects the Court of Justice of the EU. *International Organization 70*(2), 377–408.

Larsson, Olof , Daniel Naurin, Mattias Derlén, and Johan Lindholm (2017). Speaking Law to Power: The Strategic Use of Precedent of the Court of Justice of the European Union. *Comparative Political Studies 50*(7), 879–907.

Lax, Jeffrey R. (2011). The New Judicial Politics of Legal Doctrine. *Annual Review of Political Science 14*(1), 131–157.

Lindholm, Johan (2019). *The Court of Arbitration for Sport and Its Jurisprudence.* T.M.C. Asser Press, The Hague.

Lupu, Yonatan and Erik Voeten (2012). Precedent in International Courts: A Network Analysis of Case Citations by the European Court of Human Rights. *British Journal of Political Science 42*(2), 413–439.

Naurin, Daniel , Per Cramér, Olof Larsson, Sara Lyons, Andreas Moberg, and Allison Östlund (2015). *The CJEU Preliminary Reference Procedures Database (1997-2008).* University of Gothenburg: Centre for European Research (CERGU).

Owens, Ryan J. , Justin Wedeking, and Patrick C. Wohlfarth (2013). How the Supreme Court Alters Opinion Language to Evade Congressional Review. *Journal of Law and Courts 1*(1), 35–59.

Sadl, Urska and Ioannis Panagis (2015). The force of eu case law: An empirical study of precedential constraint. In A. Rotolo (Ed.), *Legal Knowledge and Information Systems: JURIX 2015: The Twenty-Eighth Annual Conference*, pp. 71–81. The Hague: IOS Press.

Staton, Jeffrey K. (2006). Constitutional review and the selective promotion of case results. *American Journal of Political Science 50*(1), 98–112.

Tiller, Emerson H. and Frank B. Cross (2006). What Is Legal Doctrine? *Northwestern University Law Review 100*(1), 517–533.

Vanberg, Georg (2005). *The Politics of Constitutional Review in Germany.* Cambridge: Cambridge University Press.

Voeten, Erik (2007). The Politics of International Judicial Appointments: Evidence from the European Court of Human Rights. *International Organization 61*(October 2005), 669–701.

Voeten, Erik (2020). Populism and Backlashes against International Courts. *Perspectives on Politics 18*(2), 407–422.

Whittington, Keith E. (2003). Legislative sanctions and the strategic environment of judicial review. *International Journal of Constitutional Law 1*(3), 446–474.

Winkels, Radboud , Jelle Ruyter, and Henryk Kroese (2011). Determining Authority of Dutch Case Law. *Legal Knowledge and Information Systems 235*, 103–112.

Appendices

A Model selection and validation

In this supplementary material, we discuss the collection of training data, the process of selecting the optimal classifier and our validation of classifications for paragraphs in our test set, the CJEU's judgments in preliminary reference cases concerning the free movement of goods issued between 1998 and 2011.

A.1 Training data

At the time of writing, we had access to XML-files containing the text of the 4,265 judgments the CJEU published in all preliminary reference procedures between June 1997 and December 2019.[18] Along with unique case identifiers the CJEU assigns to its decisions, each file comes with XML-tags allowing us to distinguish between paragraphs in the text of each judgment. We selected a sample of 417 judgments, making up the training and validation sets for our classifier. We then parsed these judgments into paragraphs using the **xml2** package for the programming language R, yielding a total of 8,816 paragraphs.[19]

A.2 Hand-coding of paragraph classes in the training data

We distinguish between paragraphs belonging to one of four issue types: (1) the admissibility of the national court's questions referred for a preliminary reference (*admissibility*), (2) the CJEU's jurisdiction to address the national court's questions (*jurisdiction*), (3) preliminary observations the CJEU made prior to addressing the national court's questions (*preliminary observations*), and (4) the CJEU's reasoning on the national court's questions (*question*). Per issue type, we further distinguish between classes of paragraphs that begin the court's discussion of the issue and paragraphs concluding the issue (e.g. `admissibility_start` and `admissibility_stop`). For the issue type *question* we add a third paragraph class `question_noanswer`, reserved for paragraphs in

[18] The texts of these judgments and their metadata were scraped from the CJEU's website, Curia, and made available for this project by Stein-Arne Brekke.

[19] XML-tags for judgment texts allowed us to distinguish between different sections of judgment texts, e.g. their preamble, grounds and operative parts. We focused only on the grounds of judgments, the section of interest to us, which reduced the number of paragraphs a hand-coder had to process in the second step of our approach.

which the CJEU asserts that a national court's question does not require an answer. While these paragraphs in effect conclude an issue, they typically stand alone in a judgment's text and are semantically different from paragraphs actually offering an answer to a national court's question. All other paragraphs which neither begin nor conclude an issue are classified as `arguments`.

A trained hand-coder then assigned one of these paragraph classes to each paragraph found in the 417 judgments selected for our training and validation set. Prior to training the classifier, we removed a list of common stop words from each paragraph's text and stemmed the remaining words. We found that relying on 3-grams rather than individual words yields better classification results and therefore tokenize words into 3-grams. Finally, we removed 3-grams which only appeared in a single paragraph.

We randomly sampled 100 of the 417 judgments with hand-coded paragraphs for our validation set, while paragraphs from the remaining 317 judgments served as training data. Table 3 displays the frequency distributions of paragraph classes in the training and validation sets. Table 3 indicates that some paragraph classes appear very infrequently in our data, particularly `jurisdiction_start` and `jurisdiction_stop` with only four paragraphs each (and none of them being sampled into our validation set). For these paragraph classes, it is unlikely that there is sufficient information in the training data for the classifier to adequately classify paragraphs in the validation set.

Further, the modal paragraph class in our data is `arguments`, with far more paragraphs than the second and third most frequent classes, `question_start` and `question_stop`. This suggests that any classifier is likely to achieve relatively high accuracy rates without necessarily making meaningful predictions in the context of our approach. The classifier could simply predict the class `arguments` for every paragraph in our validation set, yielding a high proportion of accurately predicted paragraphs. However, we require a classifier that not only achieves high accuracy in general but does so for each paragraph class we are interested in, given the correct identification of paragraphs beginning and concluding an issue is the primary motivation of our approach. For a meaningful evaluation of the classifier's performance, it is therefore necessary to consider the precision and recall rates for each paragraph class we are interested in.

Table 3: Distribution of paragraph classes in training and validation set

Issue type	Paragraph class	Training set		Validation set	
		Freq.	*Prop.*	*Freq.*	*Prop.*
admissibility	`admissibility_start`	24	.004	5	.002
	`admissibility_stop`	24	.004	5	.002
jurisdiction	`jurisdiction_start`	4	.001	0	.000
	`jurisdiction_stop`	4	.001	0	.000
preliminary observations	`preliminary_start`	52	.008	20	.009
	`preliminary_stop`	34	.005	14	.006
question	`question_start`	388	.059	132	.059
	`question_stop`	386	.059	132	.059
	`question_noanswer`	19	.003	6	.003
	`arguments`	5657	.858	1910	.858
Total		6592	1.000	2224	1.000

Note: Frequency distributions of hand-coded paragraph classes from 417 CJEU preliminary reference judgments selected for training and validation set (N8,816).

A.3 Model selection

In our approach we opt for a neural network, a deep-learning classifier suitable for text classification, learning patterns in the supplied training data through an iterative process. The required data input to train the classifier is a $n \times m$ paragraph-feature matrices, with n indicating the number of paragraphs and m indicating the number of features (more specifically words or n-grams) in the training and validation sets respectively, along with the paragraphs' corresponding classes.[20]

We build our neural network using the **keras** and **tensorflow** packages for R. We find that a neural network with two layers of 480 hidden units and an output layer with 10 units (given our 10 paragraph classes) yields adequate results while remaining computationally feasible. The input layer uses a 'relu' activation function while the output layer uses a 'softmax' activation returning a probability distribution over paragraph classes for each paragraph in the validation set. We fit models for three different batch sizes across 10 epochs. Figure 7 plots the models' loss and accuracy. Plots in the top row of Figure 7 show that all three models achieve high accuracy rates above .95, which is unsurprising given the distribution of paragraph classes. The bottom row of Figure 7

[20] After tokenizing each paragraph into 3-grams, the paragraph-feature matrices are of dimensions 6592×55289 for the training set and a 2224×55289 for the validation set.

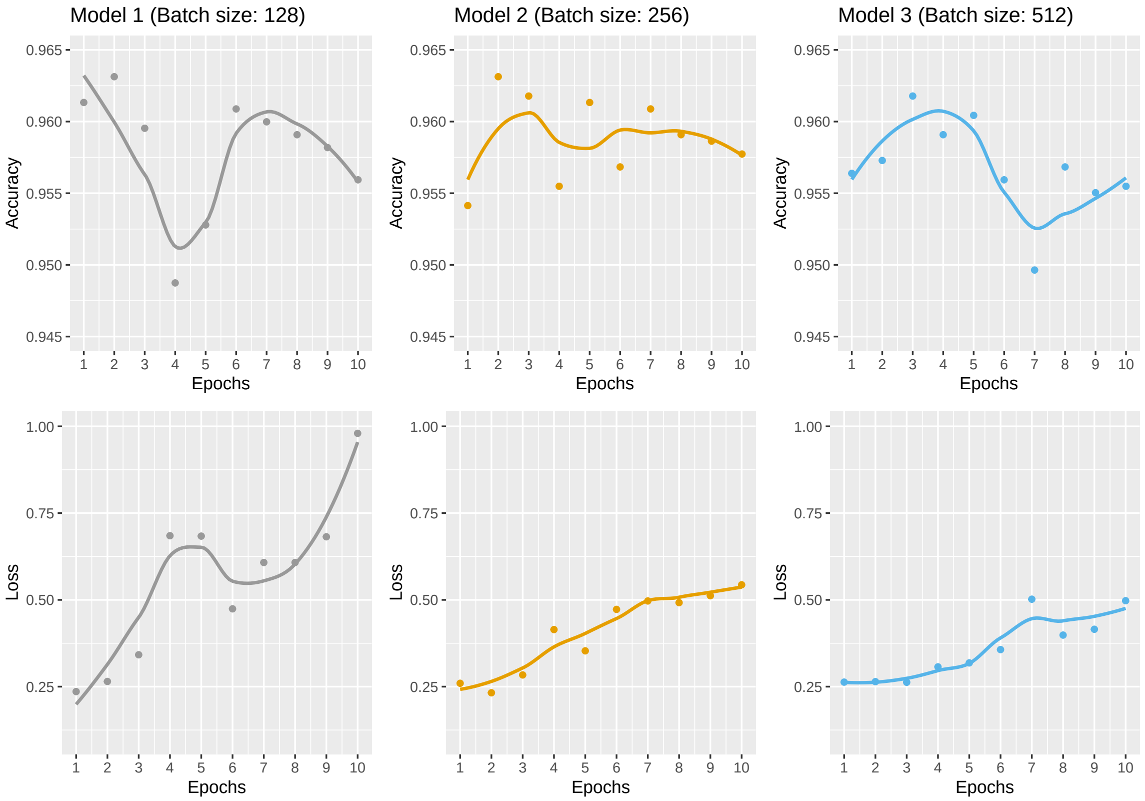

Figure 7: Model loss and accuracy for the validation set for neural networks fitted for three different batch sizes. The best-performing model minimizes loss and maximizes accuracy.

shows that all three models are prone to overfit the data as the number of epochs training the neural network increases, with only Model 2 returning a reduction in loss in the second epoch.

Table 4 reports precision and recall rates for the classes `question_start`, `question_stop`, `question_noanswer` and `arguments` for the three fitted models at the second epoch.[21] Table 4 shows that all three models return very few false-positive classifications (i.e. show high precision rates) across the four paragraph classes, with lower albeit still relatively high recall rates. Classification problems tend to face a trade-off between precision and recall rates. It is worth noting that for the application discussed here configurations of a neural network which would yield higher recall rates at the expense of precision would be available. We prioritized precision over recall as this facilitated the validation of classifications in the test set, further discussed below.

[21] Note that given the low frequency of the remaining paragraph classes in the training data, we focus our attention on these four classes.

Table 4: Precision and recall rates for fitted neural network models

Paragraph class	Model 1		Model 2		Model 3	
	Batch size: 128		Batch size: 256		Batch size: 512	
	Precision	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
`question_start`	0.94	0.82	0.95	0.80	0.95	0.73
`question_stop`	0.97	0.89	0.98	0.86	0.99	0.84
`question_noanswer`	1.00	0.86	1.00	0.86	1.00	0.86
`arguments`	0.97	0.99	0.96	0.99	0.96	0.99

Note: Precision refers to the proportion of predictions (per class) which were predicted correctly by the model, i.e. true positives/(true positives + false positives). Recall refers to the proportion of actual instances of a class the model predicted correctly, i.e. true positives/(true positives + false negatives).

A.4 Validation and improving classification performance

In light of the classifier's performance on CJEU decisions in the validation set discussed in the previous section, we have reason to expect that predictions are correct for a majority of paragraphs contained in the test set decisions. However, for the purpose of our application presented above we require that all paragraphs in the test set are correctly classified.

For the validation of the classifier's predictions for the test set we built an online interface allowing coders to view the text of a decision's paragraphs along with their predicted classes.[22] Through the interface, coders can quickly spot any of the classifier's mistakes or omissions and correct them. Therefore, our approach promises to significantly reduce the time it takes a human coder to process and validate (predicted) paragraph classes per decision, as few paragraphs require an active intervention by a human coder.

Further, validated paragraph classifications in the test set can be used as additional training data for future applications of using a supervised machine learning classifier to split court decisions into issues. As validation progresses, the continually growing number of validated paragraphs available as training data for our classifier promises to reduce errors in future predictions. We demonstrate this benefit of our approach by comparing the performances of a classifier across different volumes of training data. At the time of writing, we had access to a total of 19,731 validated paragraphs, 8,816 paragraphs from our original training data and 10,915 paragraphs from our test data. We split the 19,731 paragraphs into training sets of different volumes (4,000, 8,000 and 12,955 paragraphs) and a

[22] Further information and demonstrations of the interface are available on request.

Table 5: Comparison of classification performances for different volumes of training data

Paragraph class	Training data: 4,000 paragraphs		Training data: 8,000 paragraphs		Training data: 12,955 paragraphs	
	Precision	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
`question_start`	0.87	0.78	0.94	0.70	0.80	0.85
`question_stop`	0.88	0.78	0.94	0.78	0.89	0.89
`question_noanswer`	1.00	0.33	0.93	0.75	0.89	0.70
`arguments`	0.98	0.99	0.97	0.99	0.99	0.99

Note: Precision and recall rates for paragraph class predictions of a neural network with two hidden layers comprising 480 nodes each.

validation set, comprising 6,775 paragraphs. Table 5 shows precision and recall rates for paragraph classes `question_start`, `question_stop`, `question_noanswer` and `arguments` for a neural network trained on 4,000, 8,000 and 12,955 paragraphs respectively. Table 5 shows that recall rates improve as the volume of training data increases, albeit to some extent at the expense of the classification precision. Nonetheless, we ultimately strive for a classifier providing both high recall and precision rates, and the classifier trained on the largest volume of training data comes closest to achieving this goal.

B Network analysis

In this section we describe in greater detail the network or graph analysis elements of the study.

B.1 Generating the network

We began by identifying from the text of the studied judgments (source) all references to other, previous CJEU judgments (target), including both judgments that are part of the studied sample and those which are not. We recorded in which paragraph of the referring judgment the reference is found and, where available, the referred to paragraph of the target judgment. The sample contains 2,476 such references. We treat those references as edges in a time-directed network.

Take for example the following excerpt from in Case C-478/07, *Budějovický Budvar, národní podnik v. Rudolf Ammersin GmbH*, ECLI:EU:C:2009:521, para. 98:

> It follows that, since the bilateral instruments at issue now concern two Member States,

Table 6: Three edge versions

	Source	Target
Paragraph-level edge	ECLI:EU:C:2009:521–98	ECLI:EU:C:2003:295-37
Issue-level edge	ECLI:EU:C:2009:521:I3	ECLI:EU:C:2003:295:I1
Judgment-level edge	ECLI:EU:C:2009:521	ECLI:EU:C:2003:295

> their provisions cannot apply in the relations between those States if they are found to be contrary to the rules of the Treaty, in particular the rules on the free movement of goods (see, to that effect, Case C-469/00 *Ravil* [2003] ECR I-5053, paragraph 37 and the case-law cited).

As presented in Table 6, the references from *Budvar* to *Ravil* can be represented as an edge in three different ways that are all accurate but differ in terms of specificity. The paragraph-level edge is the most specific. By discarding the paragraph information we can easily create judgment-level edges. However, we can also use the paragraph information to identify the source and target issue respectively which, in turn, allows us to create issue-level edges of intermediate specificity. Thus, using the same information we are able to create case law citation networks of varying degree of specificity: a paragraph-level network (not part of the study), an issue-level network, and a judgment-level network.[23]

B.2 Network centrality

We then calculate centrality for all vertices in both the issue-level and judgment-level network. We seek to capture the persuasiveness of the Court's reasoning and, more specifically, how well embedded it is in existing jurisprudence. There are a few centrality measurements worth considering for this purpose, each with certain advantages and disadvantages.

The most simple and straight-forward measurement is *outdegree* which is equal to the number of out-going references. The main advantages with outdegree is that it does not change[24] and that it is calculated locally, i.e. it is independent of the rest of the network and therefore also of sampling. Them main drawback with outdegree is that it (in its basic form) attaches equal weight

[23] This means that only judgments and issues that contain references are included in the networks and are capable of receiving centrality scores.

[24] The number of references in a judgment are and always will remain the same regardless of how case law develops.

to all edges. In the context of this study it means that it only captures the quantity of references without regard to quality, which is non-ideal.

For this reason, outdegree is frequently replaced or supplemented by *hub score.* Hub score, which is one side of the HITS-algorithm, was developed for the purpose of identifying web pages that link to good authorities (Kleinberg 1999). Like outdegree, a vertex's hub score is based on its outward edges, but instead of only reflecting how many other vertices a vector is connected to hub score also reflects how many other vertices point to those target vertices. Thus, hub score incorporates a qualitative element not present in outdegree. However, hub score tends to perform less well on small data sets and is sensitive to small changes.

We here consider if and to what extent the results differ if instead of outdegree we use hub score or hub rank, a variant calculated on the basis of target vertices page rank rather than authority (see Derlén and Lindholm 2017).

C Additional results

In the following, we present additional results from our regression analyses, replacing our outcome variable *Outdegree* with the alternative measures for network centrality discussed in the previous section, *Hub Score* and *Hub Rank.* For each outcome variable, we again estimate three regressions, (1) a judgment-level regression, (2) a pooled issue-level regression, ignoring variation over judgments beyond the control variables included in the models, and (3) a partial-pooling multi-level regression allowing intercepts to vary across judgments.

Both *Hub Score* and *Hub Rank* are non-negative continuous variables, with values concentrated in the left tails of their distributions (this is particularly true for *Hub Score*). Figure 8 plots of both variables for the judgment- and issue-level. We can see that there is hardly any variation on the variable *Hub Score* for both levels (with most values concentrated at or only marginally above zero and only one value at 1), underlining concerns that calculating hub scores to measure network centrality is infeasible for relatively small datasets like ours.

Given *Hub Score* and *Hub Rank* are continuous variables with (heavily) right-skewed distributions, we opt for Generalized Linear Models with a Gamma distribution and a log link function for our regressions. To account for the zero-values on many of our observations at both the judgment

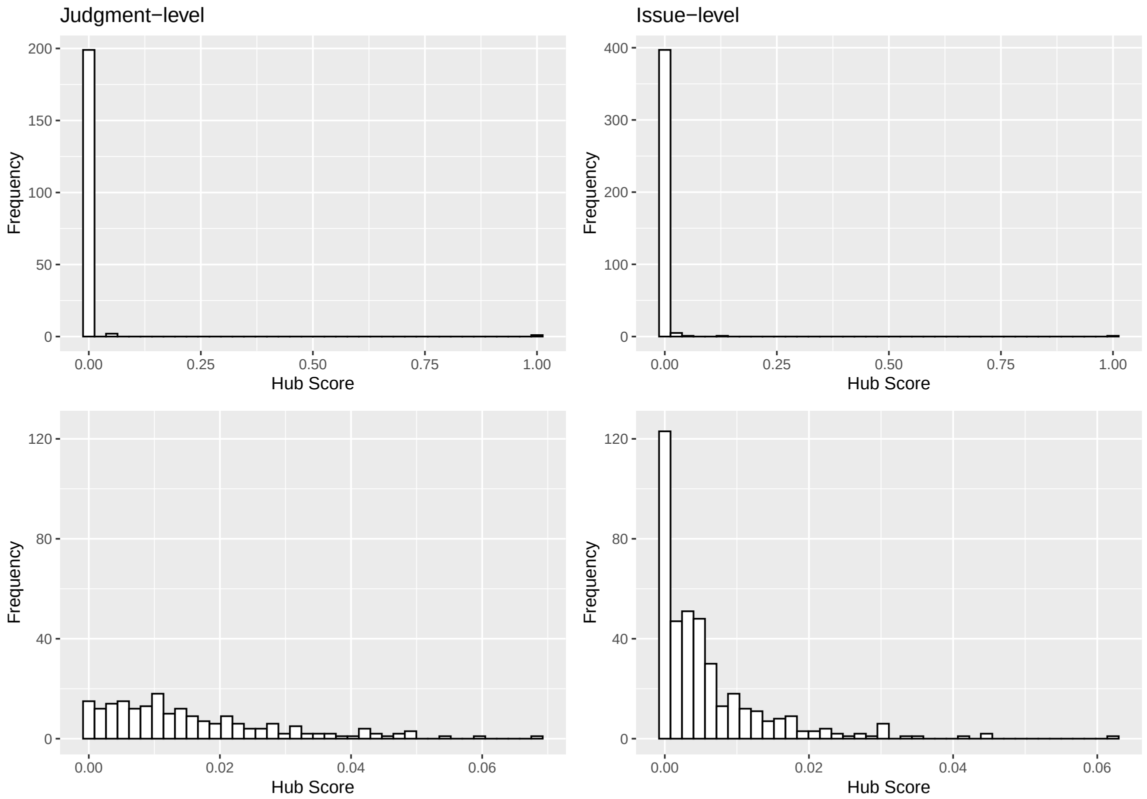

Figure 8: Distributions for outcome variables *Hub Score* and *Hub Rank* at the judgment-level (N205) and issue-level (N405).

and issue-level, we consider linear transformations of the outcome variables, adding a small value of 0.001 to each observation.

Coefficient estimates for the various regression models are displayed in Figure 9, along with their 95% HPDs. Overall, we find that coefficient estimates for models including *Hub Rank* as outcome variable are by and large similar to the results discussed in Section 4.3 of the main manuscript. Notably, coefficients for the category *In conflict* of the variable *MS Conflict* remain positive, albeit they are no longer clearly distinguishable from zero for any of the models, with the lower tails of the coefficients' posterior distributions overlapping zero. In addition, similar to the results reported in the main manuscript, coefficient estimates for the category *In conflict* of the variable *AG Conflict* are positive and distinguishable from zero, suggesting that the CJEU is more likely to cite existing case law with higher precedential authority when its position runs counter to the Advocate General's expressed position, a result that holds across the judgment- and issue-level analyses.

In contrast, results for regression models with *Hub Score* as the outcome variable differ signif-

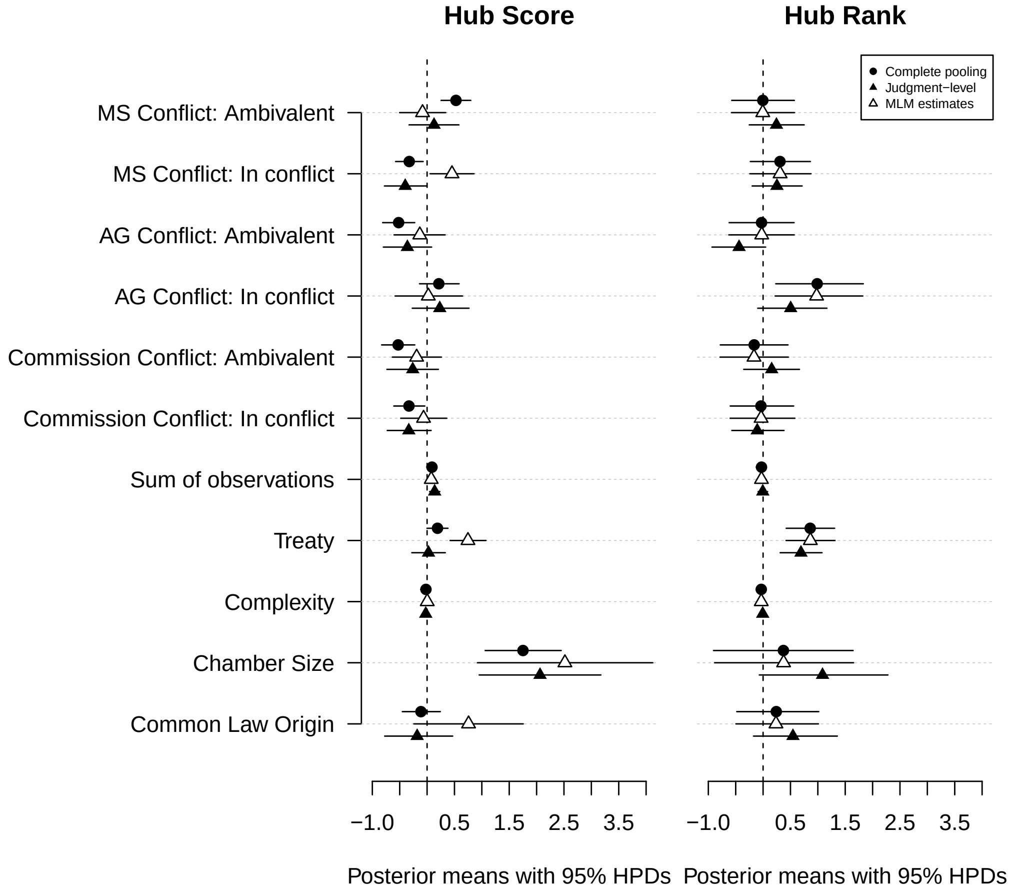

Figure 9: Posterior means with 95% HPD intervals of regression coefficients, displayed for the judgment-level (N205), issue-level and multi-level analyses (N405). All regression analyses include year fixed-effects (not shown here).

icantly from results reported in the main manuscript, although we remain cautious about these results as we have reason to expect that the estimated hub scores do not reflect a valid measure of network centrality for our data.