QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES

Lluc Font Pomarol

Quantitative large-scale analysis of judicial decisions: judicial disruption and practices

Lluc Font i Pomarol

DOCTORAL THESIS

UNIVERSITAT ROVIRA i VIRGILI

2023

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

Quantitative large-scale analysis of judicial decisions: judicial disruption and practices

Lluc Font i Pomarol

DOCTORAL THESIS

Supervised by:
Dr. Roger Guimerà Manrique
Dr. Marta Sales Pardo
Dr. Sergio Nasarre Aznar

DEPARTMENT OF CHEMICAL ENGINEERING

UNIVERSITAT ROVIRA i VIRGILI

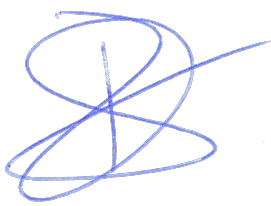
Tarragona, 2023

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

Declaration

WE STATE that the present study, entitled "Quantitative large-scale analysis of judicial decisions: judicial disruption and practices", presented by Lluc Font i Pomarol for the award of the degree of Doctor, has been carried out under our supervision at the Department of Chemical Engineering of this university, and that it fulfills all the requirements to be eligible for the International Doctorate Award.

Doctoral thesis supervisors:



Dr. Roger Guimerà Manrique Dr. Marta Sales Pardo Dr. Segio Nasarre Aznar

Tarragona, August 28th, 2023

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

Abstract

In the past 30 years, technological advances in computation have allowed to storage, process, and analyze massive amounts of data resulting from all sorts of human activities. The ability to address these large-scale data is crucial in developing new understandings of the sociological and cultural aspects underlying these human activities. In the case of legal studies, digital resources from court and legislative activities (such as legal codes and judicial decisions) can be easily accessed in public repositories. Although the legal domain does not rely in computational and quantitative approaches as much as other fields do, the use of such techniques has increased significantly over the years, with many scholars exposing the benefits of adopting empirical and quantitative methodologies to generate objective, falsifiable and reproducible knowledge. In the present thesis, we use network science and statistical inference tools over large-scale corpora of judicial decisions to reveal and understand patterns behind the functioning of the judicial system.

The data we use encompasses digitized documents corresponding to approximately 100,000 judicial decisions ruled by courts in the Spanish judicial system. These documents are structured in three different corpora including decisions from the legal domains of housing, homicides and condominium, respectively. Besides from the text written by the reporting judge, we also have access to other metadata such as the list of cited legislation and precedents, the date, the court, the names of the justices, among others. The corresponding cases are mainly ruled by courts of first appeal (*Audiencias Provinciales*) but there is also cases from the Supreme Court (*Tribunal Supremo*) and other high courts. Taking these data, we rely on network representations of the content of the data to quantify the content of our documents and be able to organize and compare them among each other. Specifically, we consider bipartite networks where the nodes from one group are the documents and the nodes from the other group are

either the words in the text or the legislation cited in such documents. This network representation will allow us to tackle different aspects of the judicial system.

The first aspect we study involves quantifying the evolution of content of judicial decisions over time, finding trends and shifts and detecting periods in which disruptive topics arise. To do so we first quantify the content of documents using a network-science-based topic model and, second, we measure the time variation by means of information-theoretic metrics. Since our document quantification goes beyond the usual word-wise representation and is also able to capture the legislation used in each document, our results offer a more detailed interpretation specially suited for legal documents. In particular, we are able to identify an abrupt change in housing-related decisions around 2016. Moreover, because our information-theoretic approach pinpoints the specific content that drives change, we are also able to interpret the results in terms of the role played by legislative changes, landmark decisions, and the influence of social movements.

The second aspect we study concerns the relation there might exist between the attributes of judges and the content of the decisions they write. Specifically, we measure the extent to which knowing the words or the legislation used is predictive of the attributes of the judges. On the one hand, we quantify the content of decisions in terms of function words, content words or the cited legislation, using the same approach as in the first aspect we study. On the other, we extract the gender, the seniority and the identity of the judge. To measure the differences there exist between these subgroups, we use a machine learning classifier that learns how to discern between them from the content of their decisions (on part of the decisions) and then validate the learned classification criteria by making predictions of the class of the judge (on the remaining decisions). The general idea is that the better the predictions are, the more important the difference between these subgroups is. Given that we are able to quantify the content of decisions in terms of either stylistic or content-related features, we are able to interpret better the nature of the differences we measure. Our results show that there are inherent differences in the way judges write and apply the law that makes them distinguishable, opening the door to predict their attributes.

Finally, the third aspect we study is the role of novel behaviors in judicial sentencing. Specifically, we measure the degree of novelty in a judicial decision in terms of the combinations of cited legislation they do. By using a network representation of the decisions and the cited legislation, we define different novelty metrics. First, a definition of novelty in terms of how atypical pairs of citations are and, second, a definition in terms of pioneering a citation strategy. By making an analogy with scientific publications, where the presence of uncommon referenced articles can be linked with scientific impact, interdisciplinary and success, we study the relation that might exist between the innovative practices in terms of cited legislation and the future impact of a judicial decision. By considering the reuse of content of decisions performed by future judges to write new decisions as a proxy for the impact of a decision, we are

able to use the degree of innovation of a decision to predict the impact it will have in the future. These results show that there exist a relation between these novel practices and the adoption of the written content of them in the future.

All in all, in this thesis we laid bare the potential behind the use of large-scale data and computational tools to study documents from courts in the context of legal studies. By using tools from network science, statistical inference and information theory, we studied different aspects behind the functioning of the judicial system. Certainly, having a deeper understanding of the judicial system is crucial to scrutinize it, to then open the door for legal scholars and policy makers to improve or emend those aspects that require to.

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

Agraïments

Com no podria ser d'una altra manera, aquest treball no hauria estat possible sense la contribució, directa o indirecta, d'un seguit de persones. És per això que els ho vull agrair.

En aquesta tesi, hem mirat d'aplicar mètodes de la ciència de dades a la recerca en dret. El fet d'haver pogut bastir un pont entre aquests dos mons, que sovint es troben tant llunyans, ha estat possible, sobretot, gràcies a l'inestimable expertesa i vocació dels meus directors de tesi: la Marta Sales, el Roger Guimerà i el Sergio Nasarre. A ells, els agraeixo l'oportunitat de participar en aquest projecte i, sobretot, tot allò que he après durant aquest temps.

També m'agradaria dedicar unes paraules d'agraïment a totes les persones que han passat pel SEESLab al llarg d'aquesta tesi. Encara que vam haver de passar per l'aïllament i el teletreball a causa de la COVID, aquest recorregut no hauria estat el mateix sense poder desconnectar una estona amb els dinars i els cafès. Gràcies, Sergio, Oscar, Ignasi, Lluís, Alejandro, Angelo, Oriol, Maribel, Teresa i Manu. També vull agrair a tots els membres de la Càtedra d'habitatge de la URV, ja que sempre s'han posat a disposició per ajudar-me i fer més amena la meva incursió al mon legal.

M'agradaria agrair també a l'Albert Díaz i a la Luce Prignano que m'obrissin les portes del mon dels sistemes complexos i la ciència de xarxes, i sense els quals segurament no hauria decidit embrancar-me en aquest projecte.

I per últim, agrair als meus pares i a la meva germana, a qui potser moltes vegades no sé explicar prou bé els detalls d'aquesta tesi, el recolzament sense

el qual tota aquesta feina (i molts altres aspectes de la vida) no haurien estat possibles. I també a la Júlia, perquè m'entén.

Contents

Abstract **v**

Agraïments **ix**

List of Figures **xv**

List of Tables **xix**

List of Abbreviations **xxi**

1 Introduction **1**

1.1 A debate on the legal method 3

1.2 Scope of the work 5

2 Data representation and methods of analysis **9**

2.1 Data representation of judicial decisions 11

2.1.1 Network science to reveal large-scale patterns in legal citations 11

2.1.2 Data representations of text 14

2.2 Judicial decisions from the Spanish judiciary 17

2.2.1 Three corpora of digitized judicial decisions 17

Reporting judge 20

2.2.2 Pre-processing the digitized text of a corpus of judicial decisions 31

2.2.3 A complex network topic model applied to judicial decisions . . . 34

3 Revealing disrupting topics and epochs from judicial decisions 41
3.1 Topics give a global view of the evolution of decision contents . 43
3.2 Bayesian surprise reveals disruptive periods and topics 47
3.3 Legal interpretation of disruptive topics 49
3.4 Conclusion 56

4 Language and the use of law to predict judge gender and seniority 61
4.1 Feature selection and judge attributes from judicial decisions . . 64
4.2 Judge attribute prediction from decision content-related features 66
4.2.1 Judge identity is highly predictable from language and use of legislation 66
Judge identity prediction is linked to content reuse . . . 68
4.2.2 Judge gender can be predicted from content-related features 71
4.2.3 Judge seniority can be predicted from content-related features 72
4.2.4 Function words are as predictive of seniority and more predictive of gender than content word 74
4.2.5 Gender and seniority differences are attributed to complex combinations of features 75
4.3 Conclusion 78

5 Atypical and pioneering law citations can predict content reuse 83
5.1 Novelty in judicial decisions based on cited legislation 84
5.1.1 Novelty as atypical combinations of cited legislation . . 85
5.1.2 Novelty as pioneering combinations of cited legislation . 88
5.1.3 Atypical novelty and group novelty characterize different aspects of innovation 90
5.1.4 A legal interpretation of innovation 91
5.2 Decision impact based on content reuse 95
5.3 Relation between novelty and impact 97
5.4 Conclusion 100

6 Conclusions and perspectives 101
6.1 Perspectives and future work 103

A Bayesian estimation of the Kullback-Leibler Divergence 107

A.1 A hierarchical Bayes point estimate for β 110

B Word and legislation topics **115**

C Random forest algorithm **123**
C.1 Definition of Random Forest 125
C.2 Performance metrics 127
C.2.1 Classification 127
C.2.2 Regression 128
C.3 Cross-validation testing 129

D List of top-novelty judicial decisions **131**

Bibliography **133**

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

List of Figures

2.1 Number of decisions per year for the three different corpora . . 18
2.2 Representing judicial decisions as bipartite networks 19
2.3 Probability and complementary cumulative distribution function for the number of citations 21
2.4 Complementary cumulative distribution function for the number of decisions per judge 22
2.5 Fraction of judicial decisions written by female reporting judges by year and corpus. 23
2.6 Fraction of decisions written by female judges by court in homicides corpus 24
2.7 Fraction of decisions written by female judges by court in condominium corpus 25
2.8 Fraction of decisions written by female judges by court in housing corpus 26
2.9 Fraction of decisions written by early-career judges by court in homicides corpus 28
2.10 Fraction of decisions written by early-career judges by court in condominium corpus 29
2.11 Fraction of decisions written by early-career judges by court in housing corpus 30
2.12 Probability distribution function for the number of words per judicial decision for the three different corpora 35
2.13 Hierarchical topic models and decisions as distributions over word topics 37

2.14 Hierarchical topic models and decisions as distributions over legislation topics . . . 38
2.15 Legislation topics law entropy at the different hierarchical levels (HL) of the model . . . 39

3.1 Time evolution of word topics and legislation topics . . . 45
3.2 Time evolution of word topics at the different hierarchical levels of the model . . . 46
3.3 Time evolution of legislation topics at the different hierarchical levels of the model . . . 47
3.4 Kullback-Leibler divergence to measure surprise in the time evolution of topics . . . 50
3.5 Kullback-Leibler divergence in the time evolution of word topics for different hierarchical levels . . . 51
3.6 Kullback-Leibler divergence in the time evolution of legislation topics at the different hierarchical levels . . . 52
3.7 Topics in housing-related decisions that change abruptly and contribute to disruption in 2016 . . . 53
3.8 Housing word topics contribution to disruption in 2016 . . . 58
3.9 Housing legislation topics contribution to disruption in 2016 . . . 59

4.1 Using the content of a judicial decision to predict the identity, the gender and the seniority of judges . . . 62
4.2 Differences in language use and law citation are highly predictive of the reporting judge . . . 67
4.3 Correlation between word reuse and legislation reuse . . . 70
4.4 Judge gender and seniority prediction . . . 73
4.5 Function words are as predictive of seniority and more predictive of gender than content words . . . 75
4.6 Judge gender prediction performance for content and function words for different information-content thresholds . . . 76
4.7 Judge seniority prediction performance for content and function words for different information-content thresholds . . . 77
4.8 Performance dependence on feature selection in gender prediction . . . 79
4.9 Performance dependence on feature selection in seniority prediction . . . 80

5.1 Cumulative distribution of atypical novelties function for different corpora 87
5.2 Block novelties cumulative distribution function for different corpora 90
5.3 Housing corpora novelty metrics pair-wise correlation 92
5.4 Homicides corpora novelty metrics pair-wise correlation 93
5.5 Condominium novelty metrics pair-wise correlation 94
5.6 Legal interpretation of innovation 95
5.7 Spearman's correlation rank between novelty metrics and impact 98
5.8 Predictive power of decision impact given novelty metrics . . . 99

C.1 Decision tree for classification and regression problems 124
C.2 Binary classification performance scheme 127

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

List of Tables

B.1 Words in topic 16, hl=3 117
B.2 Words in topic 108, hl=2 and topic 377, hl=1 118
B.3 Words in topic 1, hl=3 119
B.4 Words in topic 14, hl=2 and topic 468, hl=1 120
B.5 Law articles in topic 24, hl=2 121

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

List of Abbreviations

AP	*Audiencias Provinciales* Spanish Courts of first appeal
AUROC	Area under the receiver operating curve
CDF	Cumulative distribution function
CCDF	Complementary cumulative distribution function
CJEU	Court of Justice of the European Union
ECHR	Eurpoean Court of Human Rights
ECJ	European Court of Justice
FN	False negative
FP	False positive
HL	Hierarchical level
hSBM	Hierarchical Stochastic Block Model
IR	Information Retrieval
KL	Kullback-Leibler (divergence)
LEC	Spanish civil procedure law (*Ley de Enjuiciamiento Civil*)
LDA	Latent Dirichlet Allocation
LOPJ	Spanish Organic Law of the Judicial Power (*Ley Orgánica del Poder Judicial*)
MSE	Mean squared error
NLP	Natural Language Processing
PDF	Probability distribution function
TN	True negative
TP	True positive
TS	Spanish Supreme Court (*Tribunal Supremo*)
TSJ	Spanish Regional Supreme Courts (*Tribunales Superiores de Justicia*)
SBM	Stochastic Block Model

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

A la meva família

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

1

Introduction

In 1957, Fred Kort, from the University of Connecticut (United States), undertook the following challenge: given a set of judicial cases, he attempted to deduce a mathematical expression that could predict their verdict from the initial facts and conditions. The challenge, easier said than done, focused on cases of the United States Supreme Court related to the 'right to counsel'. In particular, Kort studied cases where criminal offenders demanded the revision of their conviction, on the basis that the right to counsel had not been guaranteed during the corresponding trial (Kort, 1957). The work by Kort was indeed well motivated: first, he had noticed the impossibility to establish a pattern that could explain the position of the Supreme Court against those cases:

> It is because of the difficulties of detecting a consistent pattern in these cases by conventional methods of qualitative appraisal that a quantitative method is used in this study.

Second, justices of the Supreme Court had explicitly stressed their opinion on the viability to perform such a quantitative task:

The due process clause is not susceptible of reduction to a mathematical formula[1].

Through the search for a mathematical model that could explain the application of the law in the mentioned cases, Kort aimed at testing the consistency of the Court. In the natural sciences, if a system is deterministic, its time evolution is uniquely determined once some initial conditions are set. Similarly, if the Supreme Court is consistent (deterministic), there should exist a mathematical expression that could link the initial conditions with the verdict, modeling how the law is applied. Here, the initial conditions are the facts and the characteristics of the case, and should the model be found, they would uniquely determine the verdict.

Kort proceeded with the following scheme: given the selection of cases, he first ordered them chronologically and divided them into a train and a test group; he would use the first group to derive the mathematical expression and the second to validate how good the expression classified unseen cases. The mathematical expression is nothing more than a heuristic to derive the numeric value for the importance of the factors intervening in the verdict. In the judicial opinions they write, judges stress the importance of these factors in a qualitative way. Therefore, Kort needed to convert these qualitative assessments into numerical values. The rules in this conversion process needed to be objective and therefore equal for each case. Among the factors included by Kort there is the gravity of the crime, personal handicaps of the defendant, procedural irregularities, etc. Taking the train group of cases, results show how the heuristic is able to separate between those where justices voted in favor of the petitioner and those where they voted against, that is, the heuristic assigned higher numerical values to the former than to the latter. When applied to the test set of cases, aiming to validate the extrapolation of the method, the same heuristics and separation rule allows classifying correctly 12 out of 14 cases.

All in all, the study by Kort proved two things: first, that the Supreme Court applied consistent criteria when deciding the cases in the study and that these criteria were independent of the specific justices appointed[2]. And second, that these criteria could be quantified, proving wrong the statement made by the justices a few years earlier.

[1] See Gibbs v. Burke, 337 U.S. 773 (1949)

[2] The cases studied by Kort spanned 24 years, during which several new justices were appointed. Thus, justices vary over the cases studied.

The methodology by Kort, based on a recollection of cases to systematically read them extracting information about a specific aspect, can be labeled as systematic content analysis. Besides legal studies, this methodology has been used in other fields as a way to summarize the state of the art in a specific research question (see, for instance, a systematic review in psychology by Agteren et al., 2021 or in medical science by Garg et al., 2008). To enable scientific progress, it is crucial to assess the current understanding of a particular scientific question, exposing the limitations of existing approaches, as well as showing the disagreement and gaps that exist. Similarly, in legal studies, this methodology has been used extensively to conduct doctrinal analysis, devoted to the compilation, synthesis and analysis of cases, statues and other legal documents, with the purpose of illustrating the 'state of the art' of specific legal questions in an exhaustive manner (Hutchinson and Duncan, 2012).

Although in a very incipient way, the seminal work by Kort embodies some of the general ideas that we put in practice in the present thesis, namely, taking the textual and qualitative content of legal documents to translate some aspects of interest into quantitative representations and model them to obtain a more general and deeper understanding. Moreover, the predictive power of theories and models validates the assumptions about the functioning of the object of study. Before entering the main work of this thesis, we will introduce in this chapter the important debate about the adoption of quantitative methods in the field of legal studies and then, we will define the scope of this work.

1.1 A debate on the legal method

When Kort published his study in 1957, it probably became one of the first legal studies embracing the use of quantitative and systematic methodologies, in a clear attempt to make claims based on evidence. Since then, many other legal scholars have conducted similar analysis in a wide variety of legal fields and mostly in the US judicial system (see Hall and Wright, 2008, for a comprehensive review on systematic content analysis in legal studies). However, despite the proliferation of such studies, the use of quantitative methods is far from being the norm in the field. When analyzing the state of the art regarding a given legal question, legal scholars normally face a universe of cases (but also legislation and regulation) that can easily escape the

human ability to read them in a reasonable amount of time. Given this scenario, Panagis et al., 2016, explain how researchers tend to avoid this problem by exploring a manageable subset of cases, sometimes reinforcing their status (landmark cases) with doubtful justification. More specifically, Baude et al., 2017, illustrate this behavior with practical examples, after analyzing hundreds of doctrinal research articles published in top law reviews. Their results show that only 1 out of 4 of the articles making some claim about the doctrine had conducted a systematic review over the universe of cases, whereas the remaining articles did not conduct any systematic review nor explicitly state the universe of cases from which they selected those analyzed.

Faced with this situation, many scholars have put the legal method under debate, advocating for a revision (Stolker, 2005; Venzke, 2015). Specifically, some authors have exposed what and what not constitutes good scholarship, with a focus on the scientific nature of legal research (Gestel and Micklitz, 2014; Hesselink, 2009) and the need for the incorporation of evidence-based, quantitative and interdisciplinary approaches (Shaffer and Ginsburg, 2012; Hall and Wright, 2008; Hutchinson and Duncan, 2012; Garcia-Teruel and Nasarre-Aznar, 2022; Vick, 2004; Domènech-Pascual, 2021; Domènech-Pascual, 2022). Despite some skepticism, the positive views on adopting a legal form of empiricism dominate. For instance, Baude et al., 2017, suggest that legal academia should create some standard for publishing that binds legal scholars to a more rigorous methodology. In the same direction, Hall and Wright, 2008, expose:

> Engaging in a legal empirical method should generate knowledge that is objective, falsifiable and reproducible.

All in all, the revisionist ideas claim that a more rigorous methodology would facilitate both the evaluation of the validity and uncertainty associated with claims stemming from doctrinal research and legal studies in general. Being transparent about the methodology reduces the probability of making false statements, but most important, it contributes to the general progress of legal research as a field too. For instance, when trying to recreate a specific study or building upon previous knowledge, researchers have to start from the beginning each time, something that could be avoided when methodology is shared within the field. Finally, they advocate for approaches that combine both quantitative methodology with the subsequent interpretative and more qualitative legal analysis (Hall and Wright, 2008; Baude et al., 2017; Šadl and Olsen, 2017; Hillyard, 2007).

For this revision in the method to be possible, the acquisition of large-scale data from legal databases of court-related documents must be guaranteed. However, empirical legal researchers struggle to obtain the most important source of information for conducting research, which puts the access and usability of legal data under debate. According to Alexander and Feizollahi, 2020, access refers to not only the possibility of accessing each single digital document, e.g., court records regarding a case, but being able to do it with 'reasonable effort and cost'. As Pah et al., 2020, illustrate, this is not possible in the case of US federal court records, as the governmental system that provides them, PACER[3], applies a $0.10 fee per printed page, which causes large-scale access to these documents to have a prohibitive outlay. On the other hand, the concept of usability refers to the quality of the data: it is not enough to be able to access thousands of documents, they also need to be used to answer research questions. In other words, both the content of the documents and other data describing them (i.e., metadata, for instance the date, the court, a classification label, etc.) need to be sufficiently structured to be machine-readable. This requirement is also linked to the standardization of the content of documents: in order for algorithms to recognize and compare similar entities, judges and other law professionals should use the same format when citing case law, or referring to the law, for instance. Otherwise, quantitative research is at the expense of large-scale and time-consuming enterprises dedicated to entity recognition and disambiguation in legal documents. Thus, the focus of the debate on the method is not only on the ways research should be done, but also on the availability and 'democratization' of the sources of information that could allow for such a paradigm shift (Paley et al., 2021).

1.2 Scope of the work

Throughout this chapter, we have emphasized the methodology in the study by Kort as a way to exemplify how quantitative and systematic methodologies can be helpful for enhancing doctrinal research in legal studies. However, besides the doctrinal perspective, focused on understanding how cases are decided and, in general, how the law is applied by judges, there is also an important research domain that approaches the legal field from a social-science perspective. The judicial system can be thought as a social system

[3]Public Access to Court Electronic Records, https://pacer.uscourts.gov/

where different actors interact with each other but also with the rest of society. From this perspective, in the present thesis, we aim at using statistical and data science tools to describe and understand how the judicial system works as a whole. By using large-scale databases of court documents, we are able to answer questions that could not have been addressed by only reading a group of legal documents, but only by large-scale analysis.

First, we start by introducing, in chapter 2, the basic source of information used in this thesis. We describe the data set, the characteristics of the documents in it, as well as the techniques used to transform these documents into structured data. Specifically, the data we use encompasses written judicial decisions from the Spanish Judicial system, from which we have not only the text but also other metadata such as the identity of the reporting judge the date or the legislation cited in the text, among others. The decisions in the text include cases from three big legal areas: Housing issues, homicides and condominium, encompassing approximately 100,000 decisions. To our knowledge, we are the first to have analyzed systematically corpora of legal documents in the Spanish judicial system of the mentioned size. After introducing the data set, we illustrate the techniques used to, first, process the documents and, second, convert them in a quantitative representation by means of network science and natural language process methods. These methods will allow the main work conducted in chapters 3, 4 and 5.

In chapter 3, we use the data and document representation in chapter 2 to quantitatively track trends and shifts in the evolution of large corpora of judicial decisions, and detect periods in which disruptive topics arise by means of information-theoretic metrics. Since our document quantification goes beyond the usual word-wise representation and is also able to capture the legislation used in each document, our results offer a more detailed interpretation specially suited for legal documents. In particular, we are able to identify an abrupt change in housing-related decisions around 2016. Moreover, because our information-theoretic approach pinpoints the specific content that drives change, we are also able to interpret the results in terms of the role played by legislative changes, landmark decisions, and the influence of social movements.

In chapter 4, we also use the data representation in chapter 2, but this time to reveal how the differences in the profile of the judge affect the content of the decisions they write. On the one hand, we have a quantitative representation of the content of decisions, accounting for the words and legislation

used, and, on the other, the subgroup of the judge, namely the gender, the seniority or the identity. To measure the differences there exist between these subgroups, we use a machine learning classifier that learns how to discern between them from the content of their decisions (on part of the decisions) and then validate the learned classification criteria by making predictions of the class of the judge (on the remaining decisions). The general idea is that the better the predictions are, the more important the difference between these subgroups is. Given that we are able to quantify the content of decisions in terms of either stylistic or content-related features, we are able to interpret better the nature of the differences we measure. Our results show that there are inherent differences in the way judges write and apply the law that makes them distinguishable, opening the door to predict their attributes.

In chapter 5, we use a network science representation of the use of legislation in judicial decisions to then conceptualize novel behaviors as atypical and pioneering combinations of law articles appearing in them. By making an analogy with scientific publications, where the presence of uncommon referenced articles can be linked with scientific impact, interdisciplinary and success, we study the relation that might exist between the innovative practices in terms of cited legislation and the future impact of a judicial decision. By considering the reuse of content of decisions performed by future judges to write new decisions as a proxy for the impact of a decision, we are able to use the degree of innovation of a decision to predict the impact it will have in the future. These results show that there exist a relation between these novel practices and the adoption of the written content of them in the future.

In chapter 6, we conclude the thesis by giving final remarks and discussing future directions of the work.

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

2

Data representation and methods of analysis

Most of the documents resulting from court activity are nowadays digitized and stored in digital repositories. Typically, these digital repositories feature search engines allowing for case filtering based on the date, the judge, the legal domain, the court or the jurisdiction, among many others. In some countries, these digital repositories make cases publicly available, but in some others they can only be accessed through a paywall. While these resources are mostly dedicated to legal practitioners to ease their jobs, legal scholars also use them as a source of information to conduct their research. For them, the possibility of accessing the enormous collection of resources that these repositories offer is a necessary condition to carry out large-scale analysis. However, access is not sufficient. As it is known in information science, the simple ability to accumulate large collections of data or the ability to access them, does not directly lead to any knowledge related to the activities that might have produced these data. So to enable this knowledge, researchers must engage in conversion from raw, voluminous data, that encompass objective

facts and observations but lack context and structure, into information. This conversion requires a set of tools involving the filtration, aggregation and computation of statistics of the data, with the aim of making them useful to answer specific research questions (Ackoff, 1989; Rowley, 2007).

In the course of this thesis, we will see this conversion process from data to information applied to several legal aspects. One of them, requires dealing with legislation citations in judicial decisions, which can be extracted from each document for further analysis. Once extracted, these data is of no use until we group them by the code or statue cited, sort them by the year of the decision, separate them by the legal field of the decision or compute statistics over their use, just to name a few. These processes convert *data* into *information* and leave it ready for the posterior legal analysis in which, for instance, statistics of the use of certain legislation could be liked to changes in the law made by policymakers, thus providing some insight on a specific aspect of the functioning of the judicial system, that is, some *knowledge*. It is worth noting that devising interpretative and explainable methods of analysis along this process is a fundamental requisite in the process of transforming *data* to *knowledge* (Rudin, 2019).

Typically, when analyzing data, scholars aim at answering questions or understanding certain aspects that are complex and most times difficult to quantify. For this reason, it is common to encode the target aspect in some proxy that enables its study at scale. A proxy, in this context, should be a latent, quantifiable trace underlying this target aspect. We can find many examples of this process in a variety of fields. For instance, in the scope of 'science of science', where quantitative methodologies are used to study science as a social phenomenon, counts of citations between papers are used in multiple ways as proxies for multiple aspects behind the production of science. The most popular one is probably scientific impact. Thus, although scientific impact of a certain work might be a concept accepting multiple definitions and conveying several dimensions, it is often quantified by simply counting the number of papers that cite the work (or similar metrics built upon this measure). Certainly, counting citations is a lot simpler than reading hundreds of papers and assessing the degree of impact that a certain scholar has in each paper, besides being a lot easier to compute at scale. Similarly, the concept of 'interdisciplinarity', the extent to which certain work combines and produces knowledge in the frontier of different fields, can be quantified as the extent to which papers cite work coming from these different fields. Since science is a cumulative endeavor where new papers build knowledge upon

other previously published papers, the action of citing other papers is, in part, an acknowledgment of this previous work, where the present paper builds upon old knowledge. Thus, this encoding of the concept of 'interdisciplinarity' involves, again, the simple act of counting citations. These two examples illustrate how simple, out-of-context data, such as citations between papers, can be translated into a proxy for different, complicated aspects by simple metrics based on statistics of these citations. These examples, again, describe the process from *data* to *knowledge*.

In this chapter, we will show different ways to make similar encoding of different aspects of the judicial system, with a focus on the data set of judicial decisions from the Spanish judicial system that we have used in Chapters 3, 4 and 5.

2.1 Data representation of judicial decisions

In the digital documents supporting the judgment of a case, we can find from the most concrete items, such as the names of judges, the parties, the legal representatives, a fine given, prison sentences, etc., to the most abstract ones, such as the exposition of facts or the legal reasoning behind the final verdict. All this content is suitable to be represented in a quantitative manner, thus creating proxies that allow for an understanding of those aspects underlying the functioning of the judicial system. In the following sections, we will focus on two primary elements in the contents of legal documents, the citations between documents and the text.

2.1.1 Network science to reveal large-scale patterns in legal citations

Citations are ubiquitous in legal documents; we can find them in judicial decisions when judges make references to other cases, but also when they make references to laws, codes and statutes, or even patents. We can also find them within the legislation; codes, statues and constitutions cite each other's sections, but they also have an intrinsic interconnected structure: there are plenty of cross-references within documents, from one part of the text to another.

In judicial decisions, citing previous cases has mostly an authoritative purpose: previous cases are cited to expose how similar cases have to be decided in similar terms (but also the opposite: citing a case to show that is different enough for the decision to not follow it), following the principal of *stare decisis*

(Posner, 2000; Landes et al., 1998). These practices contribute to build the judiciary as a coherent, non-arbitrary power. This is mostly valid for common-law legal systems, such as in the United Kingdom (UK) or the United States (US), where precedent (that is, the previous cases) settle how cases have to be decided in the future, and thus the pattern of citations reveals the picture of the relevance of cases in the precedent (Fowler et al., 2007). As a matter of fact, the strong coherency in the citations of cases in judicial systems allows to easily predict citations by knowing only part of the pattern of citations (Mones et al., 2021). In other, different legal traditions, such as civil law in European countries, the role of dictating how cases have to be decided is played by legal codes instead, and thus the pattern of citations to articles and sections within these codes shows the relative weight between norms. Thus, citations are used to frame the case in the appropriated legal domain and then align it with (or dissociate it from) the relevant cases in the precedent. Nevertheless, several studies have revealed that judges also cite precedent with other, 'strategic' purposes that go beyond the mentioned ones. In international courts, such as the European Court of Human Rights (ECHR), courts are not forced to embed cases in the precedent. However, Lupu and Voeten, 2011, revealed that judges rely more on previous cases when they assume there is need for more persuasion, for instance when one of the parts is a common-law country, where this embedding in precedent is more important. Similarly, in the case of the US Supreme Court, Lupu and Fowler, 2013, revealed how judges use precedent to reinforce their decision against other, dissenting judges in the court, showing that when there is a separate opinion, that is, when there is no unanimous opinion on the decision, judges tend to make more emphasis on the precedent.

Citations in legal codes, statutes and regulations work rather differently. While in judicial decisions citations mainly have an authoritative purpose, in legal codes the purpose is informative: citations aim at incorporating, in the present text, information that is located in other sections of the same code or in other codes. Thus, these references seek to simplify the code by avoiding text repetitions, for instance. At the same time, codes have to balance the extra cost associated to looking up information in other locations. Indeed, from the point of view of the users (legislators, legal practitioners, judges, etc.), there is a complexity associated to the interconnectivity of legal codes, a complexity that goes beyond the use of language and the hierarchical structure of the code itself (Katz and Bommarito, 2014; Friedrich, 2021). This complexity can hinder those tasks associated with the acquisition of knowledge of the

text: for instance, the more interconnected a code is, the more difficult it is be for legislators to update it avoiding incoherence (Katz et al., 2020; De Lucio and Mora-Sanguinetti, 2021).

The study of citations in judicial decisions, legal codes and patents shows how citations can be used as a proxy for a variety of aspects from the judicial and the legal activity, from measuring the relevance of cases in the precedent to quantifying complexity of the legal codes. Although different in the core, these examples share a common fundamental structure: they are composed of interrelated elements that point to each other through citations. This network structure is very suitable to be represented as a graph, a mathematical object composed of vertices and edges. In the past 25 years, network science has arisen as the appropriated framework to comprehend the emerging properties of systems composed of such interrelated components, properties that can only be understood by considering not only the nature of the individual components and interactions but the pattern of connections as a whole. Taking advantage of the increase in the ability to store data and the computational power, network scientists have studied large-scale, real-world networks representing systems in a variety of domains, including social, biological, technological and informational networks. Because real-world networks are essentially non-random –links have a *meaning* and connect elements for a *purpose*– the statistical properties of them do not resemble at all those of random graphs [1]. At the same time, these networks are neither completely regular, presenting a *complex* structure that makes their description and the study of their properties challenging (Newman, 2003). Specifically, many real-word networks have a small-word behavior, which implies that while the typical shortest paths between nodes is similar than it would be expected if the links in the network were placed at random, the clustering coefficient[2] is much lower (Watts and Strogatz, 1998). Many of these networks also present a scale-free structure resulting in a power law distribution for the degrees of the nodes, which entails the presence of a small fraction of nodes with a lot of connections (spanning several orders of magnitude, Barabási and Albert, 1999). These fundamental properties, show that these networks are constrained to be created, restructured and grown, according to some mechanisms that are

[1] In a random graph, each pair of nodes is connected independently with some probability p, see Erdős and Rényi, 1959.

[2] The clustering coefficient in a graph is a measure of the tendency of nodes to be connected in triplets. Specifically, the clustering coefficient measures the fraction of inter-connected triplets of nodes over all possible triplets in the graph.

common and independent of the nature of the system: for instance, scale-free networks are associated with a rich-get-richer phenomenon; those nodes with a lot of connections have a higher probability than other less connected to acquire more of links when the network grows.

In legal studies, there have been several successful attempts to incorporate and adapt techniques from network science to study the judicial system and the law. Besides showing that these legal citation networks display the universal properties found in other real-world networks (Fowler et al., 2007), most of these efforts use node centrality metrics to quantify the relevance of cases within the precedent (Van Kuppevelt et al., 2020; Whalen, 2016), with examples in several international and national courts: Derlén and Lindholm, 2017, analyzed the case law of the Court of Justice of the European Union (CJEU); Lupu and Voeten, 2011, Olsen and Küçüksu, 2017, Šadl and Olsen, 2017, the European Court of Human Rights (ECHR); Tarissan and Nollez-Goldbach, 2016, the International Criminal Court, and Fowler et al., 2007 the US Supreme Court, to name some examples. Given that access to digital court records is easier in international than in national courts, we can find many more studies on the former. Finally, other than quantifying the relevance of cases in the precedent, network science techniques have been used for other purposes as well. An example can be found in work by Christensen et al., 2016 where the authors use the pattern of citations to classify the cases. This example shows how the citations can be used as a proxy for the content of the cases as well, being able to substitute the reading of several pages and thousands of words by just a few citations, which significantly reduced the complexity of the task to classify cases.

In the present thesis, we apply techniques from network science in Chapter 5, where we use community detection models to define and study the role of innovation and impact in judicial decisions.

2.1.2 Data representations of text

Text plays a central role in all aspects of legal studies; it is the source of information for the analysis, synthesis and interpretation of cases, rules, law and judicial decisions by which legal scholars conduct doctrinal research (Hutchinson and Duncan, 2012). Typically, text in judicial decisions tends to be dense and long, featuring technical vocabulary and abundant references to other cases, to the law or to other additional materials. Moreover, when writing decisions, judges may address several layers of audiences; beyond the parts and

the legal representatives, they also may address to lower instances (in an appellation case) or even to the political establishment (when the case is decided by a higher court, for instance, see Dyevre, 2021). All in all, judicial decisions are complex documents and hence, legal scholars have to rely in close-reading analyses to fully comprehend them and thus understand the legal reasoning behind each decision and the links between the facts and the current applicable law, among others. However, close-reading analyses are very time-consuming, and the task hardly scales with the volume of documents available. Thus, many legal scholars have followed the steps of researchers in computational social sciences and digital humanities, where quantitative and automated techniques are applied in fields such as history, sociology, literature or psychology. In these disciplines, where text is also a crucial source of information, the use of such techniques has enabled scholars to carry out studies that scale with the volume of data available nowadays. Thus, these methods provide the adequate tools to address research questions that only a large aggregate of documents can answer, and that to some extent will be answered as well throughout this thesis, such as temporal shifts in the content of decisions or the the dependence on other characteristics of the actors involved in the documents, such as the gender, the age, the role, the race, etc.

In computational social science and digital humanities, the process of conversion of text to 'data' is crucial. But language encodes the communication of information in a very complex, intricate way, a priori hindering all sorts of automated information processing techniques. For instance, syntactic structures are hierarchical and non-linear, and words can have multiple meanings depending on the context. Therefore, the complete translation of all the information contained in language into quantitative representation seems a daunting task. For all these reasons, computational models of language from natural language processing (NLP) and information retrieval (IR) do not aim at achieving a trustworthy characterization of language, but rather a characterization of specific aspects of it. As Grimmer and Stewart, 2013, point out, all quantitative models of language are wrong, but some of them are useful. In other words, the fact that there exist a vast variety of language models, does not imply they aim at performing the same task with alternative means, but rather focusing on different aspects of language and thus serve different research purposes. The following example illustrates very well this general idea. Although language is a very rich source of information, it also contains an important amount of redundancy, typically found in those words without any specific lexical meaning, but with a grammatical or structural function,

for instance. These words, usually called stop words or function words, are usually prepositions, articles, auxiliary verbs, connectors, etc., but can be also defined by using statistical techniques (Gerlach et al., 2019). Then, given this amount of redundancy, most NLP techniques tend to get rid of such words as a way to increase efficiency in the information extraction algorithms. On the other hand, function words tend to define the writing style and other idiosyncrasies of the author, and thus they are of utmost importance in fields such as literature or forensics (Ainsworth and Juola, 2018; Hughes et al., 2012). In summary, while some tasks will require disregarding function words, they are crucial for others, showing that it is the specific task and the domain which determines the NPL techniques to be used.

Despite the vast variety of existing models, they still aim at a basic common objective: enabling quantitative comparisons between different texts (Alschner, 2020). In political science, scholars might want to know if the author of a given text is more conservative than another one; in psychology, the interest may be to compare emotional states reflected in the texts; in literature, if the text of an unknown author belongs to one style or epoch, and in the legal studies of precedent, a legal scholar might want to know if the content of cases in a given court differs from higher courts, for instance (Livermore et al., 2016). In all these examples, quantitative comparison of text in documents is crucial. Then, to enable these quantitative comparisons, language models typically translate the content of documents into a low-dimensional space and thus compute geometric distances between documents in a corpus. In the most simple form, political science documents such as political speeches, manifestos, etc., can be placed in an ideological line between left and right wing ideologies (that is, a 1-dimensional space); but one could add a religious dimension (between secularism and religiousness, for instance) and obtain an ideological plane (that is, a 2-dimensional space), and so on. But in general, NPL techniques can characterize the content of documents in more abstract dimensions that are not established beforehand. To this end, the most common unsupervised machine learning[3] techniques include clustering, where the (hierarchical) structure of the documents is found; network analysis, where documents and words are placed in a network where the relative position depends on the content; vector space embedding, where each document and words are placed in a vector spaces and topic models, where

[3]Machine learning algorithms can be classified between supervised and unsupervised. Briefly, while supervised methods operate with labeled data (there is a ground truth which to learn from), unsupervised methods learn unobserved categories (or labels, structure, etc.) for which there is no ground truth.

documents are represented as latent topics (Evans and Aceves, 2016). In this thesis, we use a complex network-based topic model to quantify the corpora of judicial decisions and enable subsequent analysis in Chapters 3 and 4 (see section 2.2.3 below). While all these models make assumptions that considerably simplify the structure of language (for instance they disregard the other of words in the text), they have proved to be very useful in a wide variety of information retrieval tasks in different domains (Evans et al., 2011; Evans and Aceves, 2016).

2.2 Judicial decisions from the Spanish judiciary

2.2.1 Three corpora of digitized judicial decisions

The data set we used in the course of this thesis (see Chapters 3, 4 and 5) encompasses judicial decisions ruled by courts in the Spanish judicial system. These cases belong to three legal domains, forming three separated corpora of documents. In the first domain, cases are related to housing issues (H), mostly related to evictions, foreclosing procedures, abusive terms in mortgage contracts and squatting. In the second, cases are related to condominium issues (C), and in the third one, related to homicides (HO). With these three corpora, we cover very different legal domains including two important jurisdictions, namely civil jurisdiction (C) and criminal jurisdiction (HO), with housing decisions including cases from both jurisdictions. In all three corpora, cases are ruled by several different courts. In the Spanish judicial system, courts are organized in a hierarchical structure. At the lowest level, first instance and instruction courts hear cases from criminal, civil, social, administrative, commercial and penitentiary law. Above them, several levels of revision courts (for instance, appellation courts) have jurisdiction over cases in the level below, besides having original jurisdiction in matters where lower courts do not. Among them, courts of first appeal (*Audiencias Provinciales*) hear appellation cases from first instance cases within the corresponding province, regional supreme courts (*Tribunales Superiores de Justicia*) hear appellation cases from courts of first appeal in a given Autonomous Community (*Comunidad Autónoma*) and finally the Supreme Court (*Tribunal Supremo*) hears cases from all regional supreme courts with jurisdiction over the Spanish state. Although the vast majority of cases are ruled by first instance courts, the corresponding documents for the judicial decisions are rarely digitized and made publicly available. On the other hand, the decisions corresponding to cases in higher courts are systematically digitized and made publicly available in a national

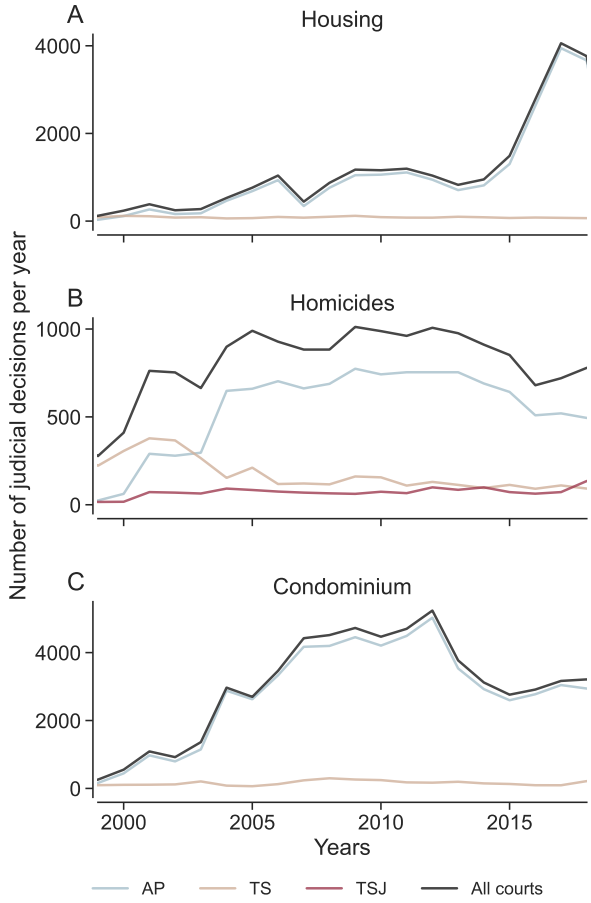


FIGURE 2.1: **Number of decisions per year for the three different corpora.** We show the number of decisions for each corpus in our data set separating by court. We only considered courts with more than 5% of decisions in each corpus. AP: courts of first appeal (*Audiencias Provinciales*), TS: Supreme Court (*Tribunal Supremo*), TSJ: regional supreme courts (*Tribunales Superiores de Jursticia*)

database[4]. For this reason, our data set encompasses mostly decisions from courts of first appeals (90%), which is the largest level of courts in terms of digitally-available cases. We also analyze cases from the Supreme Court (9%) and regional supreme courts (1%) (see Fig. 2.1 for the number of decisions in each court by corpus and year).

Although the electronic documents for these decisions are available in the

[4]The Spanish General Council of the Judiciary (*Consejo General del Poder Judicial*) makes publicly available digitized judicial decisions in the Judicial Documentation Center `https://www.poderjudicial.es/search/indexAN.jsp`

public national database, the data we used were provided to us by *Tirant Online*, one of the largest and most comprehensive databases for judicial decisions in Spain. The database includes digitized documents obtained via the public national database source published by the Spanish General Council of the Judiciary. Besides the main text, these data include, in-depth metadata that describe each document and provide relevant information obtained via parsing and entity recognition in the text. From these metadata, we used the date of the ruling, the court, the jurisdiction, the names and unified identifier of the judges or justices, the unified identifier of law articles referenced in the text, and idem for the references to cases in the case law. In Fig. 2.2A we show a scheme of the information available and extracted from judicial decisions in each corpus.

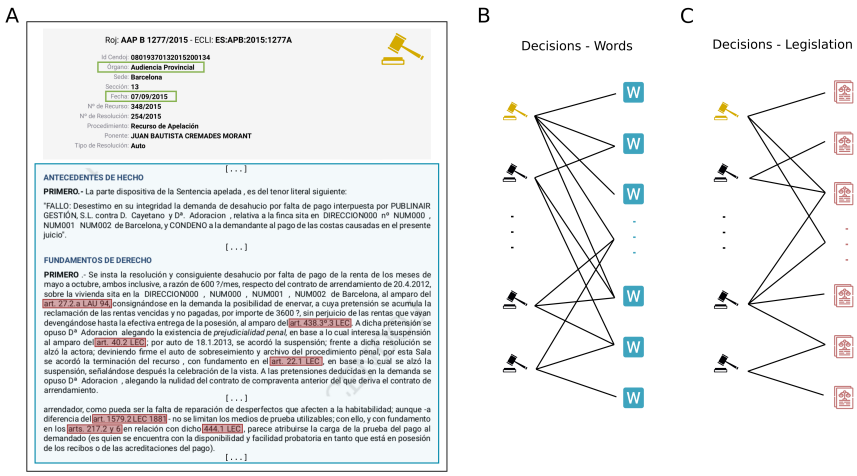


FIGURE 2.2: **Representing judicial decisions as bipartite networks.** (A) For each decision in the corpus we consider the words used (blue), the cited legislation (red), and metadata such as the date of the ruling and the court (green). We then create two bipartite networks: (B) A network of decisions and words; (C) A network of decisions and cited legislation. We select relevant words through a process of n-gram extraction and removal of words with low informational content, whereas we obtain cited legislation directly from the data.

The decisions in our three corpora span more than 100 years, with decisions back to the end of the nineteenth century and up to 2018. However, only decisions after the year 2000 have been systematically digitized by the Spanish General Council of the Judiciary, and for this reason we focus on analyzing decisions within the period 2000-2018. Given the different legal aspects that these three cases cover, their importance is different in terms of the number of

decisions they have. Thus, while there are 59,516 decisions in Condominium, there are only 22,983 in housing and 15,648 in homicides. There are also important differences in how this number varies over time, while the number of decisions is practically constant in HO, there are important variations in C and even more important in housing where the number of decisions quadruplicated from 2013 to 2016, the reasons for which we address in Chapter 3 (see Fig. 2.1). These differences also affect the length of the documents; while the mean number of words per decision is 5326 in HO, it is much lower in C and H, being 2351 and 2381, respectively (see Fig. 2.12). Regarding the number of other cases each decision cites, we observed that the vast majority of the decisions in our data set do not cite any other case (92% in H, 88% in HO and 92% in C). However, when taking a closer look in the document, we observed that citations do not follow a standardized format, and references to them are ambiguous and thus very difficult to retrieve automatically. For this reason, the data regarding citations to other cases only contains a small fraction of all actual citations. In the case of citations to the law, the data are much more reliable, since the corresponding references do not suffer from lack of standardization. While the number of citations to laws is similar in all three corpora (mean citations are 3.4 in H, 4.1 in HO and 3.6 in C), the number of articles in these laws cited is much higher in HO than in the other two corpora (mean articles cited are 7.5 in H, 13.0 in HO and 6.2 in C) probably due to a higher complexity in the codes regarding the criminal jurisdiction (see Fig. 2.3). There are important differences in the fraction of decisions that cite no legislation, being 14.5% in housing, 1.9% in homicides and 2.1% in Condominium.

Reporting judge

The courts where the decisions in our data set come from (mostly courts of first appeal and, to a lesser extent, the Supreme Court) decide cases through the consensus made by several justices in a jury. However, it is only one justice (the reporting justice) that writes and proposes the decision, and then the other justices can agree or disagree on the decision. When there is no consensus, another justice writes the decision, and the dissenting justice casts an alternative, 'dissenting' opinion (*Voto particular*, in Spanish)[5]. However, this situation is rare as justices almost always reach consensus before publishing the decision. The statistics regarding the number of decisions corresponding

[5]See articles 203-206 in the Spanish Organic Law of The Judiciary *Ley Orgánica del Poder Judicial*), https://www.boe.es/eli/es/lo/1985/07/01/6/con

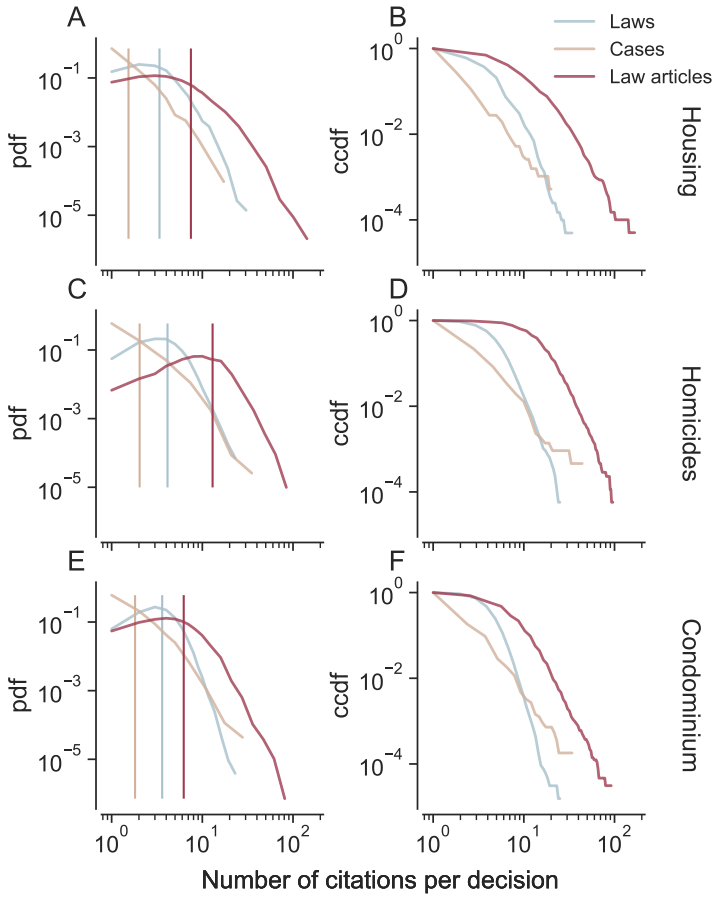


FIGURE 2.3: **Probability and complementary cumulative distribution function for the number of citations** For each corpora (row), we show the probability (left) and the complementary cumulative (right) distribution function for the number of cases, law articles, and laws cited in judicial decisions. The vertical lines indicate the mean for each each distribution, with the following values: H, 1.5 cases, 3.4 laws and 7.5 law articles; HO, 2 cases, 4.1 laws and 13 law articles; C, 1.8 cases, 3.6 laws and 6.2 law articles. The fraction of decisions that cite no legislation is 14.5% in housing, 1.9% in homicides and 2.1% in condominium.

to each reporting judge reveal important heterogeneities over the corpora. Fig. 2.4 shows the distribution of the number of decisions by justice; while the vast majority of judges has written just a few decisions, there is a reduced number of them that have written tens or even hundreds. The degree of heterogeneity is the same for all three corpora, probably because the distribution of cases among judges in courts does not depend on the legal subject.

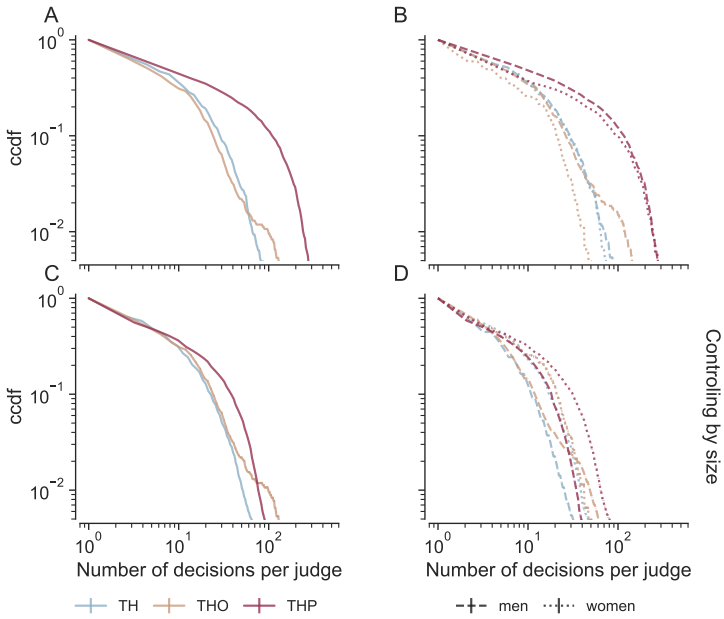


FIGURE 2.4: **Complementary cumulative distribution function for the number of decisions per judge.** (A) We show the cumulative distribution for the three corpora, housing, homicides and condominium and, (B) the equivalent separating by decisions written by a male and female reporting judge. In (C) and (D), we show the same distributions but controlling by the size of the smallest data set, which in (C) is 'homicides' and in (D) is 'homicides-women'. Results corresponding to the other cases rather that the smallest one are averaged over different realizations taking random sub-samples with size corresponding to that of the smallest data set.

The gender of the reporting judge can be inferred from the corresponding name appearing in the preamble of our documents. Thus, we divide the judges between male and female depending on whether the name appears in the list of Spanish male and female names reported by the Spanish National Statistics Institute [6]. Fig. 2.5 shows how the fraction of decisions written by female justices has increased considerably over the years, converging from a fraction of 10% to 20% in 2001, to a fraction of 35% to 40% in 2018. In the case of homicides, the fraction is systematically lower in homicides than in the other two corpora over the years. We can also find this variation over courts, with the fraction of decisions in homicides written by female justices being 1% in the Supreme Court, to 50% in Madrid Provincial Audience, for

[6]Statistics National Institute, *Instituto Nacional de Estadística*, https://www.ine.es/

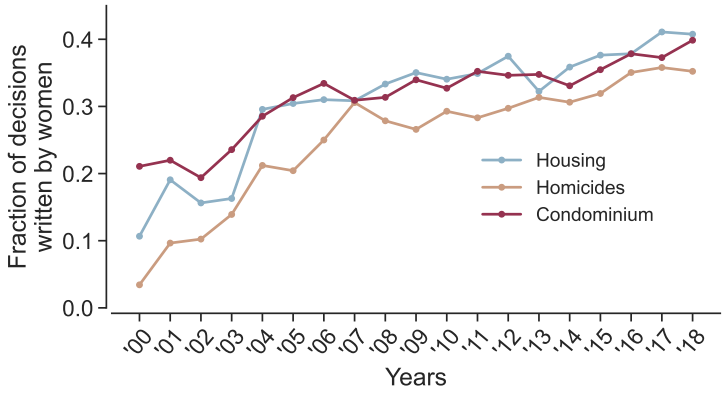


FIGURE 2.5: **Fraction of judicial decisions written by female reporting judges by year and corpus.**

instance (see Figs. 2.6, 2.7 and 2.8 for more details). Besides from these differences, there are no apparent differences in the 'production' of decisions by judge between male and female justices (see Fig. 2.4).

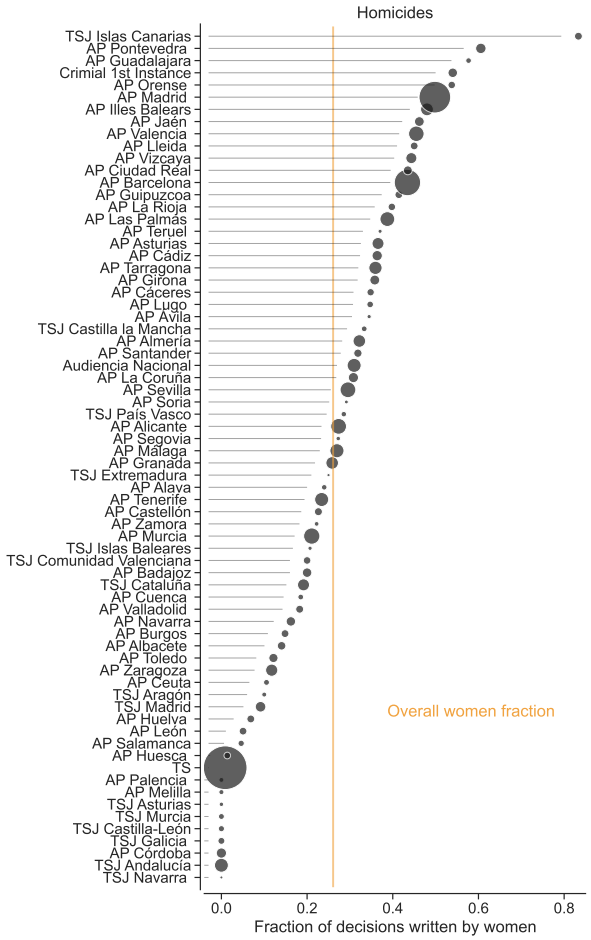


FIGURE 2.6: **Fraction of decisions written by female judges by court in homicides corpus.** The size of the circles is proportional to the number of decisions in each court in our data set. Abbreviations: TSJ: regional supreme courts (*Tribunal Superior de Justicia*), AP: Provincial Audience (*Audiencia Provincial*), TS: Supreme Court (*Tibunal Supremo*).

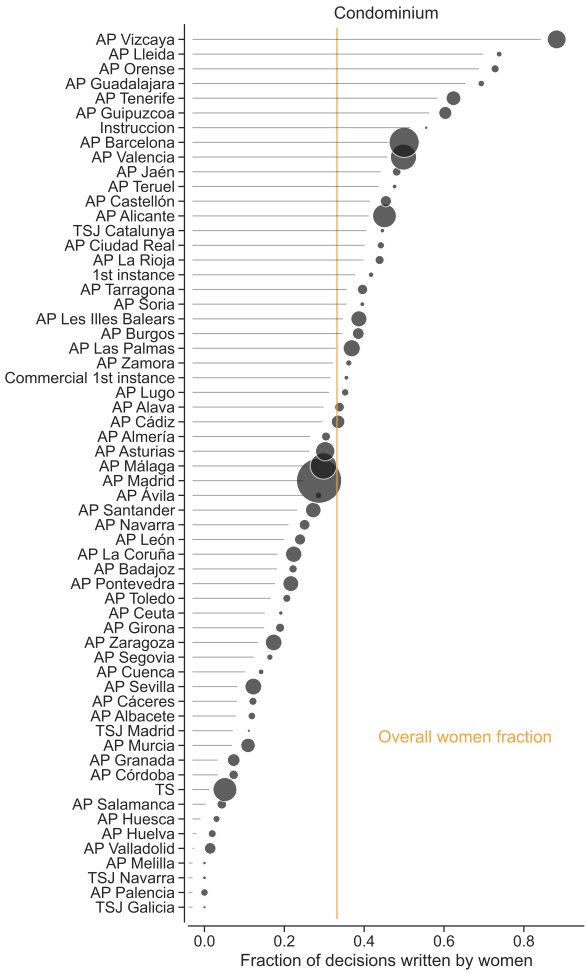


FIGURE 2.7: **Fraction of decisions written by female judges by court in condominium corpus.** The size of the circles is proportional to the number of decisions in each court in our data set. Abbreviations: TSJ: regional supreme courts (*Tribunal Superior de Justicia*), AP: Provincial Audience (*Audiencia Provincial*), TS: Supreme Court (*Tibunal Supremo*).

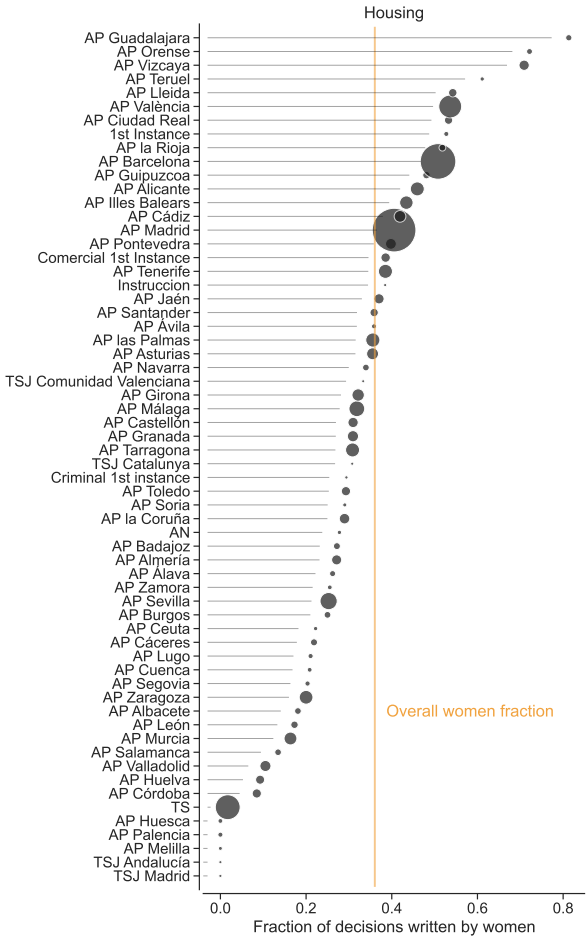


FIGURE 2.8: **Fraction of decisions written by female judges by court in housing corpus.** The size of the circles is proportional to the number of decisions in each court in our data set. Abbreviations: TSJ: regional supreme courts (*Tribunal Superior de Justicia*), AP: Provincial Audience (*Audiencia Provincial*), TS: Supreme Court (*Tibunal Supremo*).

Another characteristic of the judge that we can infer from our data set is the seniority of the judge, understood as the experience the judge in a given field. We discretize the seniority of the reporting judge by classifying them as either senior or early-career, depending on the date of their earliest judicial decision in each corpus. To avoid the effects of the change in the content of decisions over time, we restrict this analysis to decisions published in a 5-year time window, considering decisions ruled in the period 2008-2013. Moreover, to avoid having judges that have their last/first decision in the mentioned time window, we only consider those judges having both decisions ruled within the windows before 2008 and after 2013. Then, fixing the threshold at $y_{th} = 2003$, we label judges as senior those with their first decision prior to y_{th} and the opposite for senior judges. This classification results in a selection of 3,145 decisions from 375 judges in Homicides (60% of them early-career), 18,133 decisions from 435 judges in Condominium (44% of them early-career) and 3,428 decisions from 476 judges in Housing (68% of them early-career). Besides the variation over corpora, the ratio between senior and early-career judges is different over courts as well. While for homicides, the fraction of early-career decisions is 65% i courts such as the Supreme Court, it goes up to 65% in courts such as the AP Madrid (see Figs. 2.9, 2.10 and 2.11 for more details and other corpora). These differences can be attributed to differences in the promotion and retirement dynamics among courts and jurisdictions.

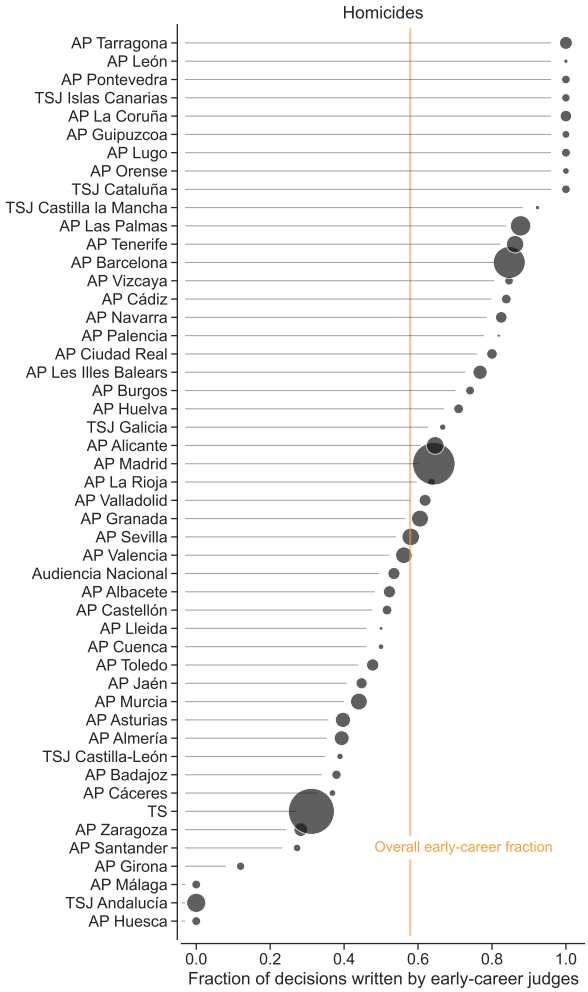


FIGURE 2.9: **Fraction of decisions written by early-career judges by court in homicides corpus.** The size of the circles is proportional to the number of decisions in each court in our data set. Abbreviations: TSJ: regional supreme courts (*Tribunal Superior de Justicia*), AP: Provincial Audience (*Audiencia Provincial*), TS: Supreme Court (*Tibunal Supremo*).

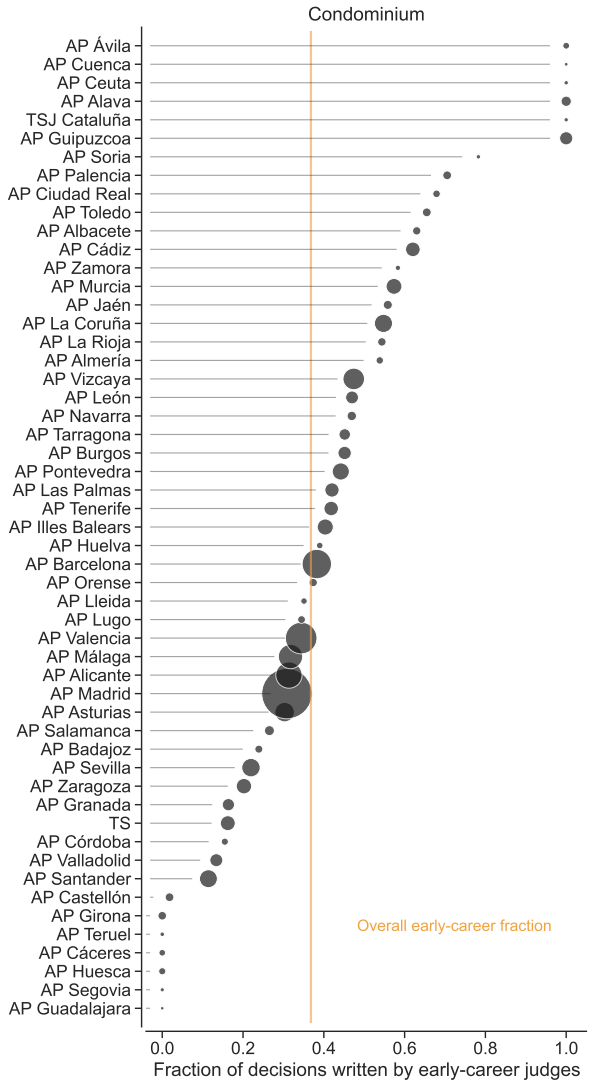


FIGURE 2.10: **Fraction of decisions written by early-career judges by court in condominium corpus.** The size of the circles is proportional to the number of decisions in each court in our data set. Abbreviations: TSJ: regional supreme courts (*Tribunal Superior de Justicia*), AP: Provincial Audience (*Audiencia Provincial*), TS: Supreme Court (*Tibunal Supremo*).

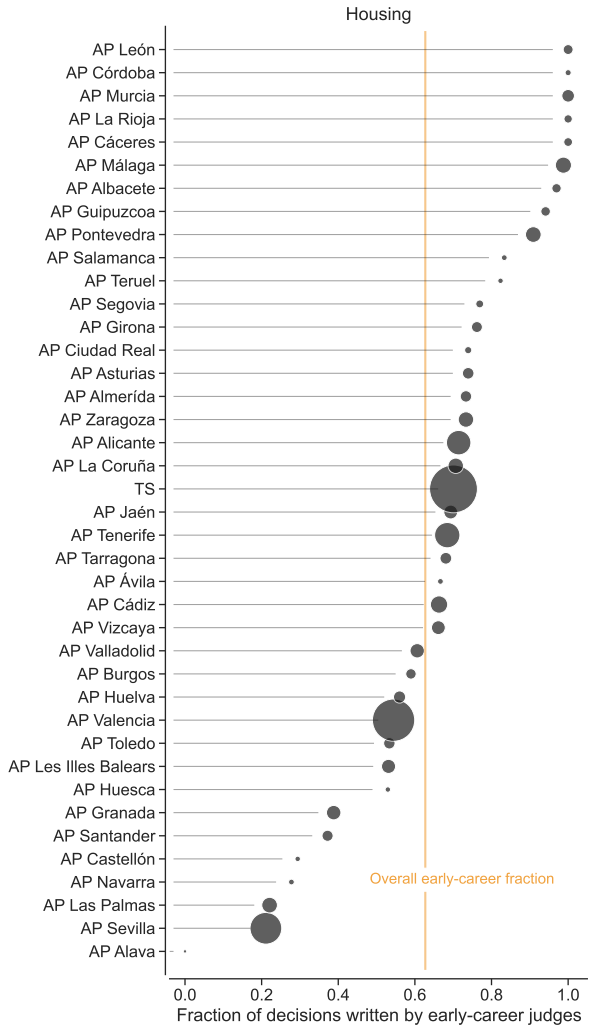


FIGURE 2.11: **Fraction of decisions written by early-career judges by court in housing corpus.** The size of the circles is proportional to the number of decisions in each court in our data set. Abbreviations: TSJ: regional supreme courts (*Tribunal Superior de Justicia*), AP: Provincial Audience (*Audiencia Provincial*), TS: Supreme Court (*Tibunal Supremo*).

2.2.2 Pre-processing the digitized text of a corpus of judicial decisions

In the previous sections, we described the data set and the fact that it contains a large collection of about 100,000 documents, each of them with a considerable length in terms of the number of words. For this reason, we need a processing protocol to clean and filter the text, enabling an efficient automated information extraction. In what follows, we describe the steps of the protocol we used, although some details and steps might vary depending on the final purpose, which we will detail in Chapters 3, 4 and 5.

Term disambiguation. When writing decisions, judges lack standardized rules to refer to common entities such as specific courts or codes. As an example, see the various ways in which the Criminal Procedure Law (*Ley de Enjuiciamiento Criminal*) can be found: 'LECrim', 'L.E.Cr.', 'L. E. Cr.', 'L.E. Criminal', 'LECRIM', 'LECrim', 'Lecrim', etc. Since those entities might be important for the subsequent legal interpretation, we homogenize then across all possible variations. We found these variations using a semi-automatic heuristic, and we replaced them by a unifying term. . Moreover, we also homogenize words by converting all characters to lower case, and we remove all numbers, non-word characters, and one-character words.

Proper name standardization and word *de-genderization.* In the text of judicial decisions there appear different proper names corresponding to the justices, parties, attorneys at law and public servants. To protect the privacy of the involved parties when decisions are made publicly available, their names are anonymized. Since we observed different ways to perform this task over the documents[7], we standardized it to avoid including more noise in the data. Then, depending on the purpose, we either substitute all names by a unique term (*'_persona_'*) or two of them: (*'José'*) form male names and (*'María'*) for female names [8]. Specifically for purposes related to gender text differences (see Chapter 4) we use the first option, and we also remove the gender declinations that correlate with the gender of the judge, for instance words such as feminine and masculine justice (*magistrado* and *magistrada* for *magistradx*).

[7]Some documents seem to anonymize names by substituting them by the term 'XXX', while some other documents replace real names for other ones preserving the gender.

[8]We perform this substitution by using a heuristic that uses the list of the most common male and female Spanish names, provided by the Spanish National Institute of Statistics https://www.ine.es/.

Beyond a bag-of-words model: extracting significant n-grams. In natural language processing techniques, it is common to assume a simplified structure of the text by disregarding the correlations that exist between consecutive words, considering the text as a list of unsorted tokens [9]. We have mitigated this assumption by considering significant n-grams, that is, chains of n tokens that appear consecutively in a corpus more than what is expected by chance. Specifically, we consider 2- and 3-grams, and we proceed as follows for a given n. We split the text (once pre-processed following the steps above) by sentences, using either periods ('.') or colons (':') as separators. Then, for each sentence, we take all possible chains of consecutive n tokens (i.e., n-grams) and, for each n-gram i we compute the number f_i of times that it appears in the corpus, and the number d_i of documents where it appears in. To select the n-grams that are statistically significant, we compare f_i with that of the null model where the order of words in each sentence is randomly shuffled, $f_i^{(nm)}$, and then compute a z-score, for each n-gram:

$$z_i = \frac{f_i - \langle f_i^{(nm)} \rangle}{\sigma_i^{(nm)}} \,. \tag{2.1}$$

Here, $\langle f_i^{(nm)} \rangle$ is the average and $\sigma_i^{(nm)}$ is the standard error of $f_i^{(nm)}$ computed over several realizations of the null model. Finally, we select those n-grams that are statistically significant by taking those with a z_i corresponding to a p-value, p, of $p < 0.05/N_{ng}$ [10]. We also apply a second condition by only considering those n-grams appearing in at least 1% of the documents ($d_i > 0.01$). Finally, we substitute the statistically significant n-grams in text: we joint the corresponding consecutive words using underscores.

Function and content word filtering using information theoretical metrics. Disregarding words that carry very little or no lexical meaning is indispensable for a successful representation of text as data. These words, called stop words or function words, are commonly filtered by using default, preexisting dictionaries (Manning et al., 2008). This methodology, although very common, entails important issues due to several reasons: first, these dictionaries only exist for some languages such as English; and second, it does not take into account the possibility that some function words could depend on the domain. Both reasons are critical in our context: first, the language of our

[9] A token refers to each occurrence of a word in the text

[10] N_{ng} is the number of found n-grams. By dividing the p-value by them, we are considering the Bonferroni correction.

data set is Spanish, which despite being the second most spoken languages in the world it does not have the computational support available for English; second, our documents come from the legal domain, which is very specific, and it has its own technical language and text structure, making them very different from other more common documents. Take for instance the word 'plaintiff'(*demandante* in Spanish). While in the context of judicial decisions this word appears in almost every document and therefore carries little meaning (for instance, it cannot be used to differentiate one document from the rest), the same word in different contexts would certainly be considered more meaningful (in a collection of newspaper articles, for instance). Thus, we would want some criteria that would define function words differently in the first case than in the second. For these reasons, we use the approach by Gerlach et al., 2019 to define and filter function words. Their methodology defines function words starting from the conditional entropy of the probability of a word to appear in a document across the corpus, H_w:

$$H_w = -\sum_d p(d|w) \log p(d|w) \ , \tag{2.2}$$

where $p(d|w)$ is the fraction of occurrences of the word w in the document over the total occurrences of the word in all the corpus. The conditional entropy of the word the extent to which a word is spread out over the documents; a word appearing in exactly the same number of times in all documents will have the maximum entropy, while another one appearing in just one document will have the minimum. Then, they define the information content of the word, I_w, as the difference between the conditional entropy in a null model (NM) (where all words have been shuffled across documents) and the conditional entropy observed:

$$I_w = \langle H_w^{(NM)} \rangle - H_w \tag{2.3}$$

In this way, a word that appears plenty of times but in just one document will have much more information content than another one only appearing once in just one document: while both will have the same conditional entropy, the first one will 'distribute' much more its probability over the documents in the null model than the second one. Gerlach et al., 2019, show that their methodology can eliminate up to an 80% of function words while still being able

to obtain a reliable information extraction by using NLP techniques such as topic modeling. Besides, the procedure is language and domain-dependent.

After following the mentioned pre-processing steps (term disambiguation, word de-genderization, n-gram extraction, and function word filtering) we end up with each of our corpus consisting of a list of documents, where each document is an unsorted list of filtered words and significant 2- and 3-grams.

2.2.3 A complex network topic model applied to judicial decisions

Once the text of the decisions in a corpus has been pre-processed, we convert it into a quantitative representation by applying a topic modeling approach. This quantification step enable the subsequent analyses described in chapters 3 and 4.

Topic modeling is an approach used mostly to classify large textual corpora and to quantify content differences between documents, among other applications, by breaking down the content of each document into latent topics, which are groups of words used similarly in documents. Blei et al., 2003, first introduced the methodology based on the general idea that the statistics of words across documents can reveal how semantically close they are, something that allows us to group words into topics. Specifically, they introduced the Latent Dirichlet Allocation (LDA), a generative model that relies on in two basic assumptions: first, that a document is generated by a mixture (distribution) of latent topics; and second, that each topic is characterized by a distribution over words. Thus, the probability of a word appearing in a document depends first, on the topic distribution of the document and then, on the probability of the word in each topic. Inferring the parameters for these distributions requires estimating the maximum likelihood given the data (statistics of words in the corpus) and considering a Dirichlet hyperprior. Although LDA has been one of the most used topic models (Blei et al., 2010), it has some important drawbacks. Specifically, Lancichinetti et al., 2015, showed how LDA has a resolution limit when it comes to detecting topics present in a small fraction of documents. This, in practical terms, comes from the roughness of the likelihood function, which implies that very different models could fit the data almost equally well and even wrong models can have higher likelihoods than the correct one (see Lancichinetti et al., 2015, for more details). Moreover, Gerlach et al., 2019, showed that the use of Dirichlet priors is not compatible with real word statistics observed in text,

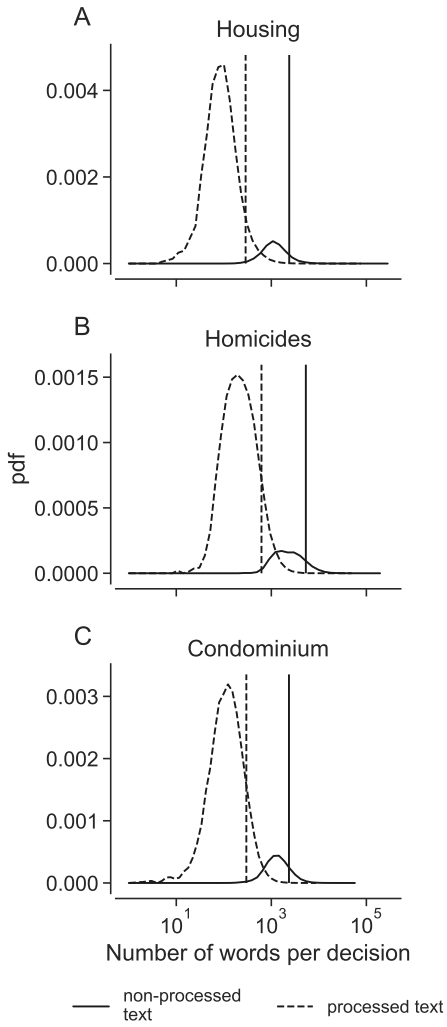


FIGURE 2.12: **Probability distribution function for the number of words per judicial decision for the three different corpora** We show the probability distribution function of the number of words per judicial decision for housing, homicides and condominium decisions. We include the corresponding distributions considering first, non-processed text and second, processed text following the steps in section 2.2.2. This processing steps include term disambiguation, number removal, one-character word removal, inclusion of significant 2- and 3-grams and function word filtering. Vertical lines indicate the mean of each distribution, with the following values: H, 2381 words (non-processed) and 288 words (processed); HO, 5326 words (non-processed) and 621 words (processed); C, 2351 words (non-processed) and 298 words (processed).

and that LDA is not able to capture data generated using hyper-distributions other than Dirichlet. Besides, Gerlach et al., 2019, showed that the problem of inferring topics is equivalent to that of inferring communities in the network formed by documents and words (see Fig. 2.2). Then, by relying on a non-parametric inference of a hierarchical Stochastic Block Model (hSBM) (Peixoto, 2014; Peixoto, 2019), they are able to overcome the limitations of LDA, while using fewer assumptions on the underlying structure of the data. In probabilistic terms, the inferred hSBM is the most plausible one given the data; in information-theoretical terms, it has the shortest description length (Rissanen, 1978), that is, it is the model that best compresses the observed data. Moreover, the model is hierarchical, in the sense that it also models the structure of the topics at higher levels, that is, modeling topics of topics at successive different levels.

Taking the approach by Gerlach et al., 2018, we are able to describe each document in our corpus as a distribution over the topics inferred. The model also provides the details of each topic in terms of the words that belong to them. Because topics are organized hierarchically, documents can be modeled with different degrees of coarse graining. On the one hand, lower levels in the hierarchy tend to describe very specific concepts. On the other, we find very general topics at higher levels, containing words that are only vaguely related. We illustrate these details in Fig. 2.13. Making an analogy with the appearance of words in documents, we also model the appearance of legislation, that is, the references to law articles made by judges in the text. If a topic model is thought to reduce the dimensionality of the representation of text, now we seek to do the same for legislation, going from thousands of possible cited articles to a few hundreds or even tens of legislation topics. By taking the bipartite network of documents and law articles (Fig. 2.2C), we also obtain hierarchically-nested legislation topics. Thus, if word topics tend to group words that can be used in similar contexts (see Fig. 2.13), legislation topics group law articles that are similar to some extent (see Fig. 2.14). In Appendix B we show the complete details in terms of words and legislation of a selection of topics, as well as their hierarchical structure.

We can illustrate this similarity by the extent to which legislation topics tend to group articles belonging to the same law, statute or constitution. We evaluate this tendency as the entropy of the distribution of laws in each topic:

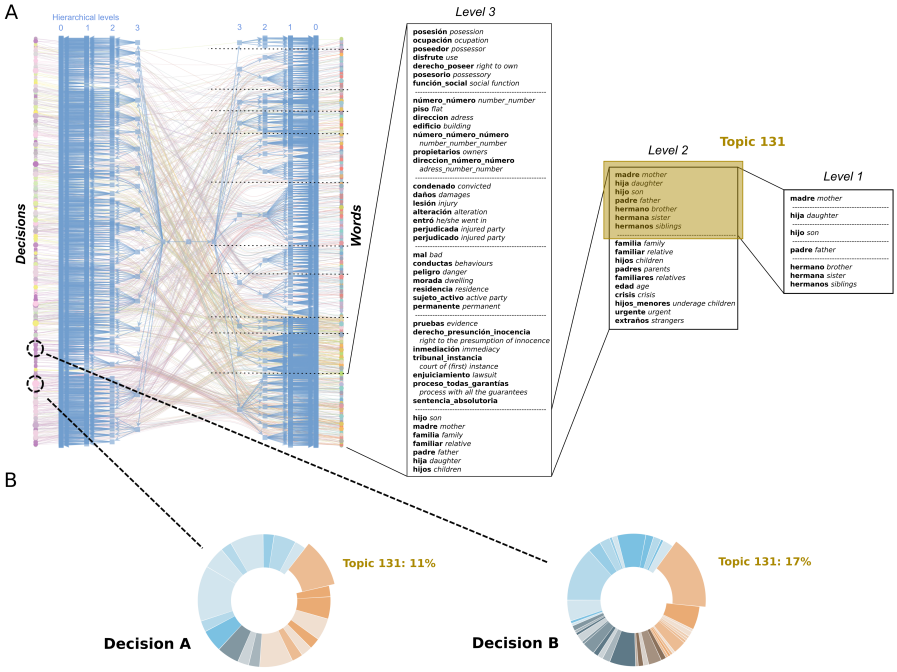


FIGURE 2.13: **Hierarchical topic models and decisions as distributions over word topics.** (A) Using a network-based topic modeling approach that uses hierarchical stochastic block models (Gerlach et al., 2018), we infer hierarchical partitions in the bipartite network of judicial decisions and words. We show word topics in the housing decisions corpus. The hierarchical structure of the model is illustrated by expanding a particular topic: from left to right, we take a topic at the highest level of the hierarchy (level 3) and expand specific sub-topics at successively lower levels. (At level 3, only the top words in each subtopic are shown.) Topics contain words that are used in similar contexts, and this similarity increases as we go down in the hierarchy. (B) Given the membership of each word in a given topic, decisions can be characterized as distributions over topics by measuring the number of times each word appears in a document. For example, topic 131 accounts for 11% of the words in Decision A, and for 17% of the words in Decision B.

$$H(t_i) = \sum_{l=0}^{L_{t_i}} p_l \log p_l \ , \tag{2.4}$$

where t_i is a given legislation topic, L_{t_i} is the number of different laws appearing in topic t_i and p_l is the fraction of articles belonging the l law. Specifically,

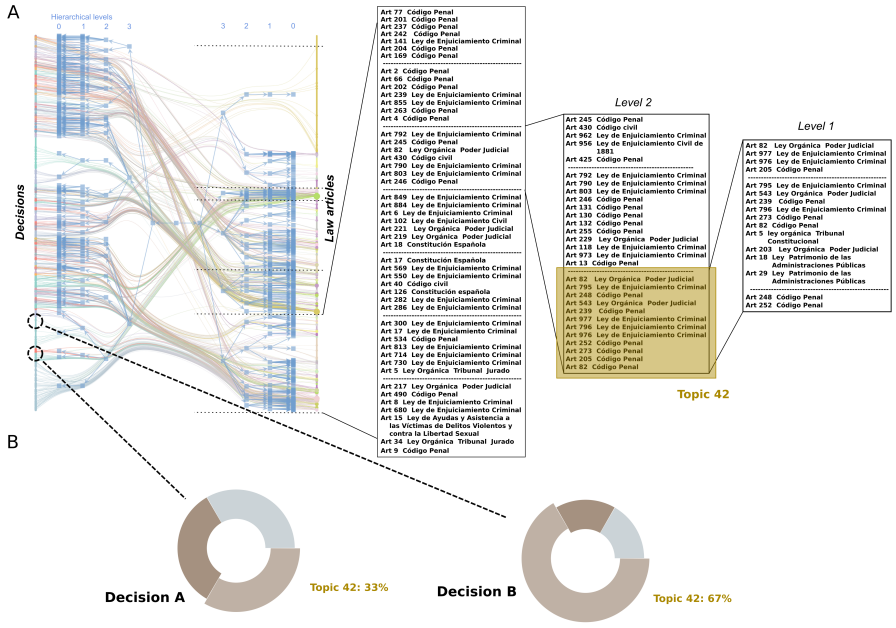


FIGURE 2.14: **Hierarchical topic models and decisions as distributions over topics.** (A) Using a network-based topic modeling approach based on hierarchical stochastic block models (Gerlach et al., 2018), we infer hierarchical partitions in the bipartite network of decisions and law articles cited in them. We show legislation topics in the housing decisions corpus. The hierarchical structure of the model is illustrated by expanding a particular topic: from left to right, we take a topic at the highest level of the hierarchy (level 3) and expand specific sub-topics at successively lower levels. We only show the top law articles in some sub-topics for simplicity.) Topics contain law articles that are used in similar contexts, and this similarity increases as we go down in the hierarchy. (B) Given the membership of each law article in a given topic, decisions can be characterized as distributions over topics by measuring the number of times each law article appears in a document. For example, taking the hierarchical level 2, topic 42 accounts for 67% of the law articles cited in Decision A, and for 33% of the words in Decision B.

we measure the significance of our results by comparing the mean law entropy (averaged over all topics) to the law entropy expected by chance, that is, a null model where law articles are shuffled around topics while keeping the sizes of the topics. We show the results in Fig. 2.15 where we observe that topics typically contain articles of the same law for all hierarchical levels of the model.

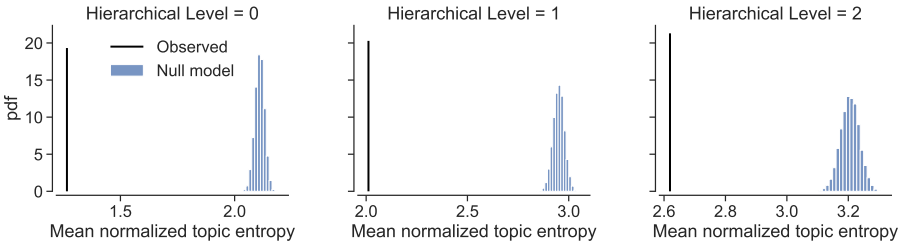


FIGURE 2.15: **Legislation topics law entropy at the different hierarchical levels (HL) of the model**. Law entropy measures the diversity in terms of laws found in the same legislation topic. We show he averaged normalized entropy over the topics and the distribution of the same quantity over the different configurations of the null model, where we shuffle law articles around topics while keeping the size of them. Data corresponds to the housing corpus

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

3

Revealing disrupting topics and epochs from judicial decisions

Technological advances have facilitated the generation and storage of digital documents stemming from human activities, from financial transactions to medical records or drug prescriptions. These digital traces open the door to understanding human behavior in new ways (Watts, 2007); digital documents allow to analyze the temporal evolution of the interactions between social actors, making it possible to infer the sociological and cultural processes beneath human activities (Evans and Aceves, 2016). In this endeavor, computational efforts are critical to automatize the processing and extraction of information from large-scale corpora of documents (Evans and Aceves, 2016). Indeed, computational methods enable the quantitative analysis of the content of documents and their evolution; they are powerful tools to understand the underlying social processes by capturing trends and patterns that result from the prevalence, extinction or substitution of specific practices and ideas (Evans and Aceves, 2016; García-Gavilanes et al., 2017).

One of the last domains to enter the digitization era is that of legal studies. In recent years, despite the fact that there are still some barriers that prevent open access to digital court records (Pah et al., 2020), there has been a steady increase in the availability of digital documents related to court activities (such as judicial decisions or processes, especially in the US and Europe) and legal processes in general (Quemy and Wrembel, 2020). In fact, despite some reluctance to incorporate evidence-based methodologies in the study of legal processes (Hutchinson and Duncan, 2012; Panagis et al., 2016; Baude et al., 2017), some voices have advocated for systematic and quantitative approaches for which the availability of digital legal documents is crucial (Pah et al., 2020; Baude et al., 2017; Hall and Wright, 2008; Gestel and Micklitz, 2014; Šadl and Olsen, 2017). Recently, systemic, computational approaches have been able to start extracting and analyzing large corpora of judicial decisions (Panagis et al., 2016; Medvedeva et al., 2020), which has enabled the study of the use and propagation of precedent through the network of citations between judicial decisions (Šadl and Olsen, 2017; Mones et al., 2021; Lupu and Voeten, 2011; Olsen and Küçüksu, 2017; Fowler et al., 2007; Charlotin, 2020). In this sense, while a thorough reading of a decision is the only way to fully comprehend the legal reasoning, computational analysis can uncover large-scale patterns in the legal system as a whole (Guimerà and Sales-Pardo, 2011; Danziger et al., 2011).

Besides enabling the systematic study of legal processes, digitized legal documents are also a good proxy for the evolution of sensitive social issues. Indeed, because of their key role in determining how societies function, law and legal decision-making are subject to public opinion (Sheshadri and Singh, 2019) and constrained to evolve and adapt to new paradigms (Katz et al., 2020; Rockmore et al., 2018; Rutherford et al., 2018). Therefore, legal documents reflect changes in culture and social norms (Klingenstein et al., 2014). Here, we investigate whether major social events leave measurable footprints in judicial records. We show that, indeed, the evolution of content in a large corpus of tens of thousands of judicial decisions from Spanish courts reveals the emergence and evolution of a socially disruptive issue. In particular, we focus on decisions related to housing in the context of the global financial crisis. Since 2007 and in less than a decade, more than 700,000 home-related foreclosure procedures were started (including those that do not result in a court procedure, as well as those that do, in all instances), which had a devastating effect on a significant fraction of the population in urban areas (Nasarre-Aznar and Garcia-Teruel, 2018). We use an information-theoretic

methodology to quantify the footprint that such a major social issue left in the judiciary, by tracking the main trends and shifts in the content of decisions, both in terms of the full text of the decisions and of their citations to existing legislation. Specifically, our analysis shows an abrupt change in the content of housing-related decisions culminating in 2016, which is in stark contrast to the smooth evolution of two control corpora related to issues that did not produce special or more than usual social unrest during the same period. Moreover, because the approach we use pinpoints the specific content that drives change, we are able to interpret the results in terms of the role played by legislative changes, landmark decisions, and the influence of social movements.

3.1 Word and legislation topics give a global view of the evolution of decision contents

We hypothesized that the major social unrest that followed the collapse of the housing market in Spain should have left measurable footprints in legal documents, and particularly judicial decisions. Therefore, we take the corpus of housing-related decisions (as detailed in Chapter 2), section 2.2.1 and we look for these measurable footprints in terms of changes both at the level of the corpus and at the level of the topics responsible for those global changes. To that end, we will use a combination of network-inference and information-theoretic approaches. To fully calibrate the changes observed in the housing corpus, we then compare them with those found in the other two corpora, homicides and condominium (see section 2.2.1). We restricted the analysis to the period 2001 to 2018, resulting in a selection of 22,983 decisions in housing, 15,648 in homicides and 59,516 in condominium.

After pre-processing by removing numbers and non-word characters, grouping and substituting statistically significant 2- and 3-grams and filtering function words, we encode each decision as an unsorted list of content words and n-grams (see section 2.2.2 for more details). Separately, we also take the list of referenced law articles. Given these two lists, words and legislation, we quantify the content of decisions by taking a complex-network topic model approach, for each corpus and for words and legislation separately (see section 2.2.3). Since the model infers hierarchically-nested topics, we are able to express each decision as a distribution over topics for each level in the model,

being able to describe the content with different degrees of coarse grain (see Fig. 2.13 and 2.14).

The model gives the membership of each word to a given topic T_i^ℓ at a given level ℓ in the hierarchy. Then, by counting the number of times each topic appears in a decision, we obtain the distribution over topics of each decision. Analogously, we obtain the yearly distribution of topics by considering all the words in all the decisions ruled in a given year—given a non-informative prior for the parameters of a topic distribution (a uniform Dirichlet distribution), the posterior yearly distribution of topics is

$$P(T_i^\ell|y) = \frac{n_y(T_i^\ell) + 1}{N_y + K^\ell} , \tag{3.1}$$

where $n_y(T_i^\ell)$ is the number of times that a word belonging to topic T_i^ℓ appears in the decisions ruled in year y, $N_y = \sum_i^{K^\ell} n_y(T_i^\ell)$, and K^ℓ is the number of topics in a given hierarchical level ℓ.

To analyze the time evolution of each topic T_i^ℓ, we calculate its importance in a given year relative to its maximum importance across years

$$E_{T_i^\ell}(y) = \frac{1}{\max_y\{P(T_i^\ell|y)\}} P(T_i^\ell|y) . \tag{3.2}$$

In the case of legislation topics, equations 3.1 and 3.2 are equivalent, but taking law articles instead of words. Then, by representing the yearly topic distribution for both word and legislation topics, we obtain a global view of the time evolution of the content of the corpora (see Fig. 3.1). By analyzing yearly changes in the prominence of topics, we observe that, for housing-related decisions, some topics that were prominent in the first years were later replaced by others (Fig. 3.1**A-B**). For instance, a word topic associated with leases loses importance, whereas one associated with abusive mortgage clauses gains prominence.

In figure 3.1**C-H**, we show the overall evolution of the relative importance of topics for the housing corpus and our two control corpora, as well as for both word and legislation topics. Qualitatively, the importance of word topics in the homicide and condominium corpora remains very stable throughout the whole period 2001-2018. By contrast, in the housing corpus, we observe a major, systemic shift in topic importance in the period 2013-2016: some word

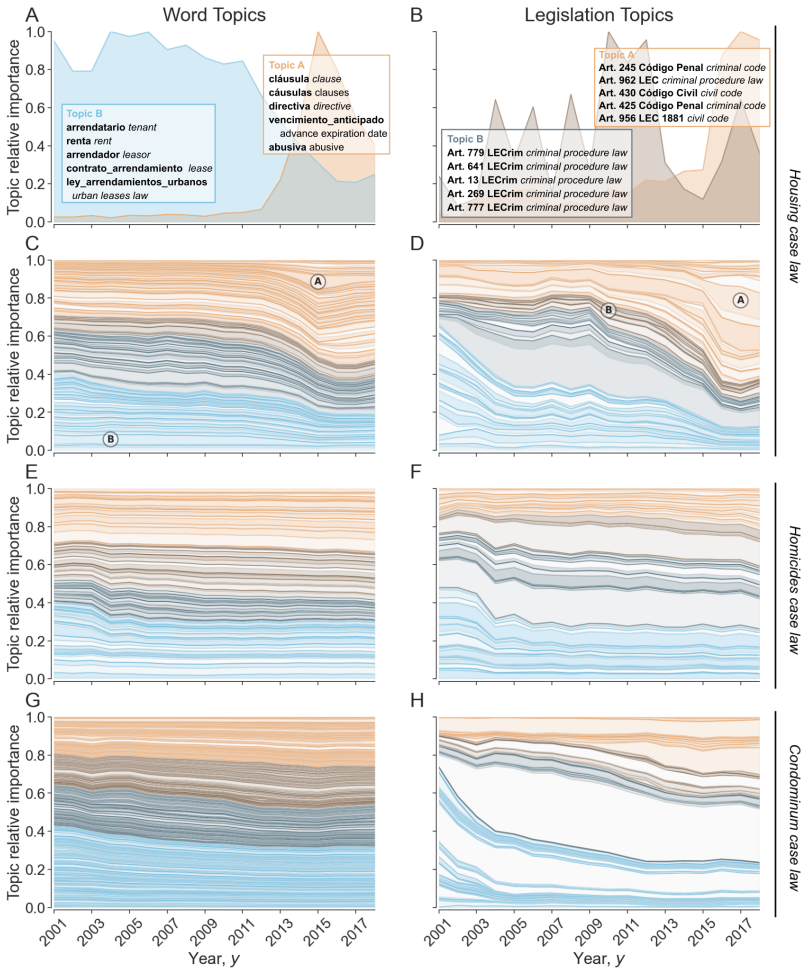


FIGURE 3.1: **Time evolution of word topics (left column) and legislation topics (right column)**. For each year y, we compute the topic distribution $P(\boldsymbol{T}|y)$ (see equation 3.1). (A-B) Time evolution of the importance of some illustrative topics in the housing corpus, normalized by their maximum importance across years (see equation 3.2). While some topics decrease in the period 2012-2018, others show the opposite behavior. Topics depicted in blue reach their maximum importance in early years, whereas orange topics reach their maximum importance in the final years. Topics that reach their maximum importance in intermediate years are depicted in gray. (C-H) Evolution of all topics for the three corpora: (C-D) housing; (E-F) homicides, and (G-H) and condominium. Each layer corresponds to a topic and their thickness corresponds to their relative importance in a given year. Topics shown in panels (A) and (B) are identified in (C) and (D), respectively. Hierarchical levels: C, L=2; D, L=1; E, L=2; F, L=0; G, L=1; H, L=0.

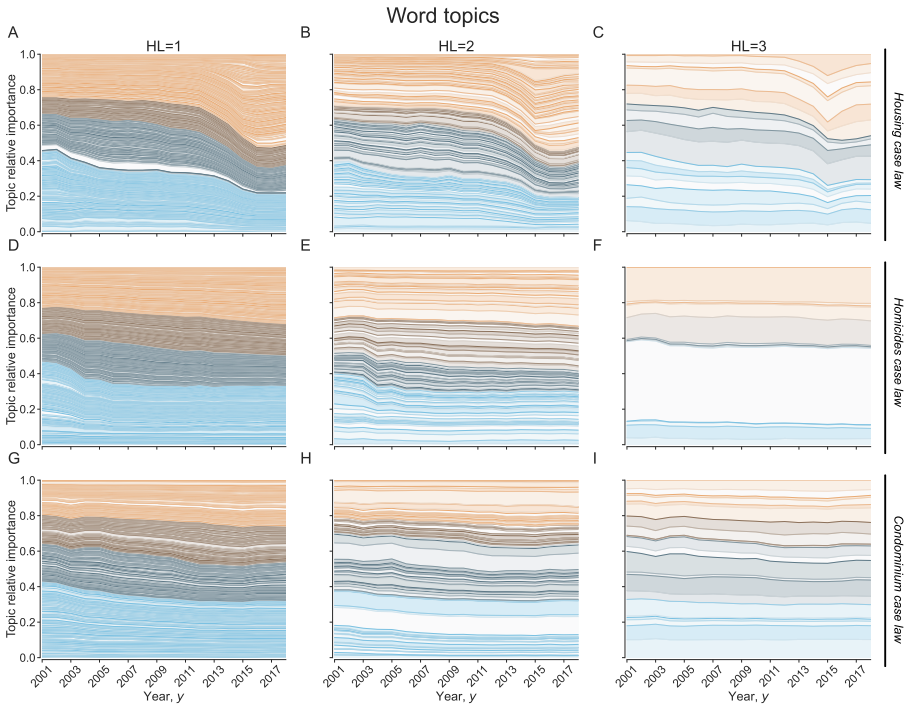


FIGURE 3.2: **Time evolution of word topics at the different hierarchical levels (HL) of the model**. For each year, we computed the corresponding topic distribution at each different HL of the model, $P(T_i^{(l)}|y_j)$ (see equation 3.1). We show the evolution of the yearly topic distribution for the three different corpora: housing (A-C), homicides (D-F) and condominium (G-I) and at the different HL of the model: from left to right: HL=1,2,3 respectively. We do not show HL=0 since it contains to many topics to be displayed. For each corpus, the qualitative time evolution is preserved through the different HL.

topics that accounted for 30% of the words in the decisions in 2011 end up accounting for over 50% only five years later. Similarly, the evolution of legislation topics is similar for the three corpora in the period 2001-2012; then a major shift occurs in housing-related decisions in the years 2013-2016, which we do not observe in the homicide and condominium corpora. Indeed, legislation topics that accounted for 20% of the cited law articles in housing-related decisions prior to 2010 end up accounting for over 60% in 2016. All these results are robust when using other hierarchical levels in the topic model (Fig. 3.2 and 3.3). Therefore, we observe a genuine shift in the language and the legislation used by judges in judicial decisions related to housing in the period 2012-2016, which we do not observe in other areas of the law.

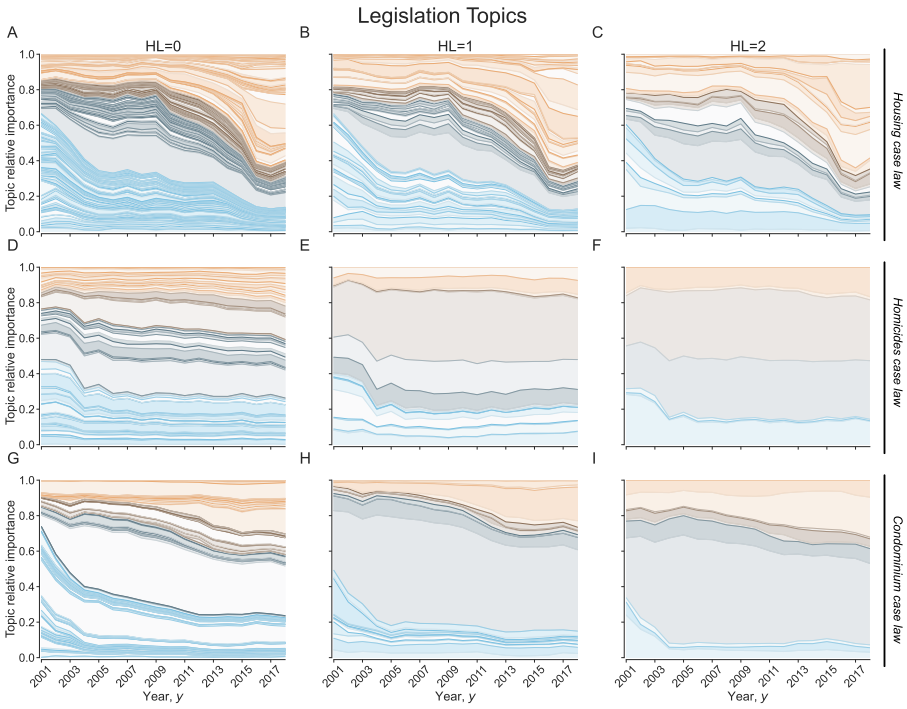


FIGURE 3.3: **Time evolution of legislation topics at the different hierarchical levels (HL) of the model**. For each year, we computed the corresponding topic distribution at each different HL of the model, $P(T_i^{(l)}|y_j)$ (see equation 3.1). We show the evolution of the yearly topic distribution for the three different corpora: housing (A-C), homicides (D-F) and condominium (G-I) and at the different HL of the model: from left to right: HL=1,2,3 respectively. For each corpus, the qualitative time evolution is preserved through the different HL.

3.2 Bayesian surprise reveals disruptive periods and topics

To quantify the extent to which the content of decisions changes, we compute the Kullback-Leibler (KL) surprise $\mathcal{S}^{\ell}_{-\tau}(y)$ between the yearly topic distribution (Eq. 3.1) at year y and that at another year in the past $y-\tau$:

$$\mathcal{S}^{\ell}_{-\tau}(y) := D_{\mathrm{KL}}(P(\boldsymbol{T}^{\ell}|y)|P(\boldsymbol{T}^{\ell}|y-\tau)) = \sum_i \mathcal{S}^{\ell}_{-\tau}(y;T_i^{\ell}), \quad (3.3)$$

$$\mathcal{S}^{\ell}_{-\tau}(y;T_i^{\ell}) := P(T_i^{\ell}|y)\log\frac{P(T_i^{\ell}|y)}{P(T_i^{\ell}|y-\tau)}\,. \quad (3.4)$$

Here, $D_{\mathrm{KL}}(\boldsymbol{p}|\boldsymbol{q}) = \sum_i p_i \log \frac{p_i}{q_i}$ is the KL divergence (Kullback and Leibler, 1951; Shlens, 2014). This metric is the relative entropy of a distribution $\boldsymbol{p}$ when expecting another distribution $\boldsymbol{q}$, and measures the average log-likelihood of observations being distributed as $\boldsymbol{p}$ when actually they were generated by $\boldsymbol{q}$ (Shlens, 2014).

The KL divergence has been shown to describe cognitive surprise (Itti and Baldi, 2009) and has been used to measure dissimilarity between speeches or texts using topic models (Rockmore et al., 2018; Barron et al., 2018; Murdock et al., 2017; Andrei and Arandjelović, 2016), word frequency models (Savoy, 2013) or other low dimensional representations of textual content (Hughes et al., 2012). Here, we adapt some of these ideas to measure the extent to which the content of judicial decisions in one year differs from those in previous years (Eq. 3.3).

To account for local temporal changes, we calculate the KL divergence between a topic distribution and the same topic distribution the year before ($\tau = 1$ in Eq. 3.3). To analyze long-term patterns, we compare them to the distribution at all previous years in the past.

Sampling factor When computing information-theoretic measures from discrete (or categorical) distributions that have been learned from data, important biases appear when the size of the sample is $N \lesssim K$, where N is the number of counts (words or law articles in our case) and K the number of categories in the discrete distribution (topics in our case). Such bias depends solely on the so-called sampling factor N/K. Thus, since our goal is to make comparisons (over the years or between corpora) of the values of the surprise, we control the sampling factor so that these comparisons are legitimate. In particular, when computing a given surprise $S_{-\tau}(y)$, we use the same sampling factor for all the years and for each corpus, so that

$$\frac{N_H}{K_H} = \frac{N_{HO}}{K_{HO}} = \frac{N_C}{K_C}\,, \tag{3.5}$$

where N is the number of words/citations used to estimate the distribution and K the number of topics for each of the three corpora. All measures reported in the manuscript are averages over sub-samples of the corpus obtained with a fixed sampling factor.

However, in Appendix A we propose a semi-analytical Bayesian estimator that makes accurate estimations of the categorical distribution in the sparse

regime and thus is able to estimate the KL with a precision that performs equally or better than existing estimators while not making *ad hoc* assumptions and being computationally less expensive. We did not apply this methodology here because we obtained the results in the present chapter before obtaining the methodology in Appendix A.

Comparing topic distributions corresponding to pairs of consecutive years (Fig. 3.4), the KL surprise quantifies the changes that are qualitatively apparent in Fig. 3.1. Until 2012, the textual content and the legislation cited in judicial decisions changes, from one year to the next, at rates that are similar among the three corpora. After 2012, the surprise between consecutive years is considerably higher in housing case law than in the two control corpora. Remarkably, both word and legislation topics display a pronounced peak in 2016, which we observe for all hierarchical levels of topics (Figs. 3.5 and 3.6). Note that measuring surprise at the lowest levels in the hierarchy reveals changes occurring in the most specific topics, while doing so at the highest level reveals changes in the most general ones. Therefore, a consistent peak throughout levels in the hierarchy implies that the changes that occurred around this date stem from a deep reorganization of topics rather than a shallow reorganization of sub-topics.

To further characterize the changes that occurred in the topic landscape, we also calculate the KL surprise between the topic distribution in one year and all years in the past. Within the corpus of housing decisions, this analysis highlights the discontinuity between decisions ruled before and after 2012 and, especially, before 2010 and after 2014 (Fig. 3.4A-B). Additionally, this analysis reveals the scope of the shift that occurred in 2016—distributions after this date show high KL divergence with respect to years earlier than 2012, but relatively low KL divergence with respect to 2016, which indicates that the shift that occurred in 2016 persists in the following years.

3.3 Legal interpretation links disruptive topics to landmark decisions and law modification

Finally, we investigate the contribution of individual topics to the KL surprise, which allows us to go beyond identifying periods of rapid change in judicial decisions, and to actually pinpoint the specific topics that most contributed to the disruption in the years around 2016. In particular, we focus on the topics that most contribute to the changes that occurred in 2016: word

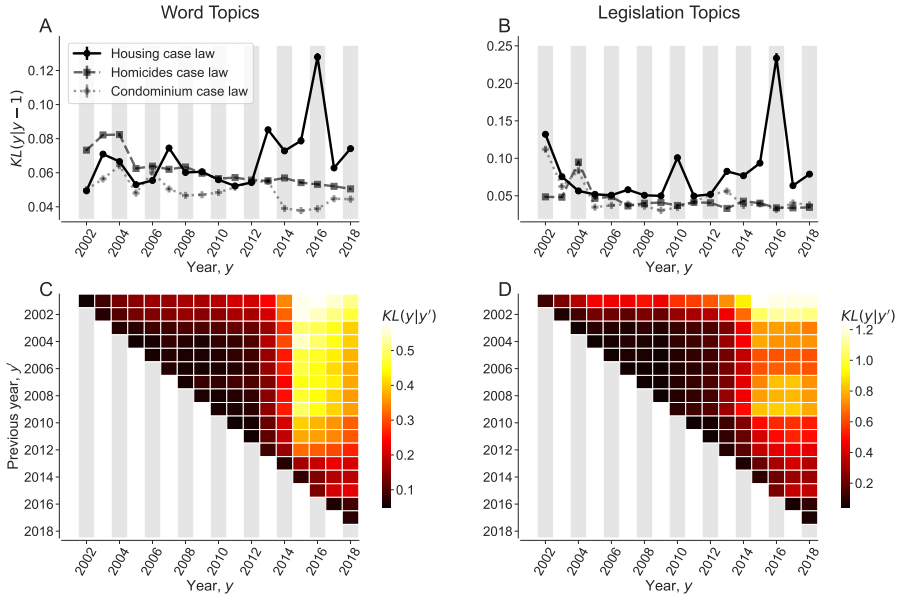


FIGURE 3.4: **Kullback-Leibler (KL) divergence to measure surprise in the time evolution of topics**. We use the KL divergence to measure the change in the topic distribution on year y with respect to the distribution on: (A-B) the year before; (C-D) a specific year y' in the past (for the housing corpus only). Although the number of decisions per year and the number of topics in the model is different for each corpus, we ensure the comparability of the results by using the same sampling factor. Error bars, often smaller than the symbols, correspond to the standard deviation of the mean over several sub-samples using a fixed sampling factor. Here, we show results for the following hierarchical levels L and the corresponding number of topics K. (A, C) Word topics for housing: L=1 and K=564 word topics; homicides: L=1 and K=688 word topics; condominium: L=1 and K=525 word topics. (B, D) Legislation topics for housing: L=2 and K=30 legislation topics; homicides: L=1 and K=17 legislation topics; condominium: L=1 and K=27 legislation topics.

topics 108 and 14, and legislation topics 41 and 42 (Fig. 3.7 A and B respectively; see also Fig. 3.8 and 3.8). Each topic legal interpretation follows the words/legislation found in the topic, some of which we include in the following (in italics; for full details on the content of each topic and their hierarchical structure, see Appendix B).

When interpreting the topics that disrupt the most, we choose the hierarchical level that balances internal coherence and the level of detail of the topics. Indeed, these two qualities behave in opposite ways when going from the

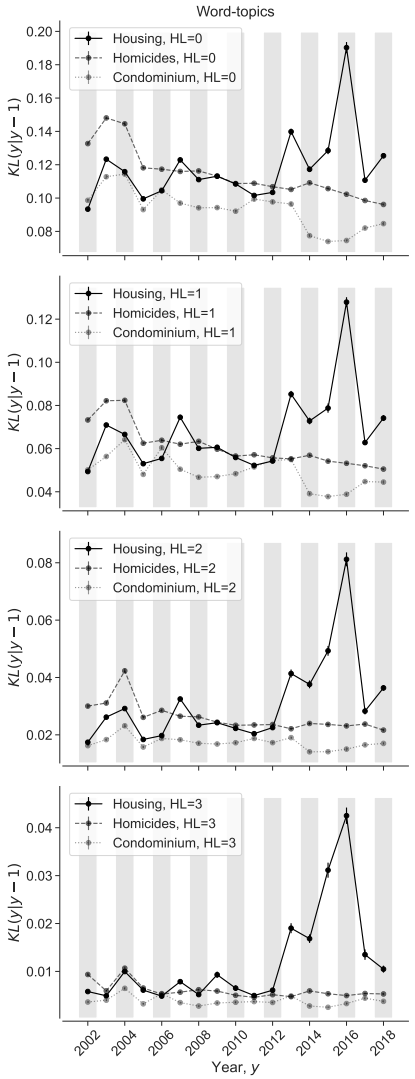


FIGURE 3.5: **Kullback-Leibler (KL) divergence in the time evolution of word topics at the different hierarchical levels (HL) of the model**.

lowest (finest) level of the topic hierarchy to the highest (broadest). For instance, if we consider word topic 14 in level 2 (see Appendix B), we observe that some sub-topics present very specific coherence but very insufficient detail. For example, in the scope of housing, the terms *denunciado, denunciante* (respondent, plaintiff) only frame the legal analysis in the area of criminal law. It is only by considering sub-topics (that is, going up in the hierarchy)

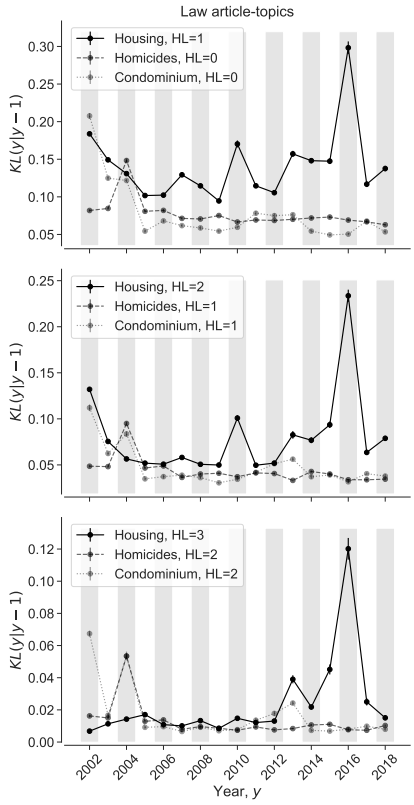


FIGURE 3.6: **Kullback-Leibler (KL) divergence in the time evolution of legislation topics at the different hierarchical levels (HL) of the model**.

that we learn more details that narrow housing criminal law to the light criminal offense of squatting. Unfortunately, going too high in the hierarchy makes interpretation more difficult because words in a topic can become specific to a set of decisions, which can distort the interpretation.

Word topic 108 Words in this topic are typically related to mortgage loan contracts, a hotly debated subject in the years after the global financial crisis of 2007, when mortgage enforcement skyrocketed. The increase in the use of such terms around 2016 stems from the following events. In the well-known

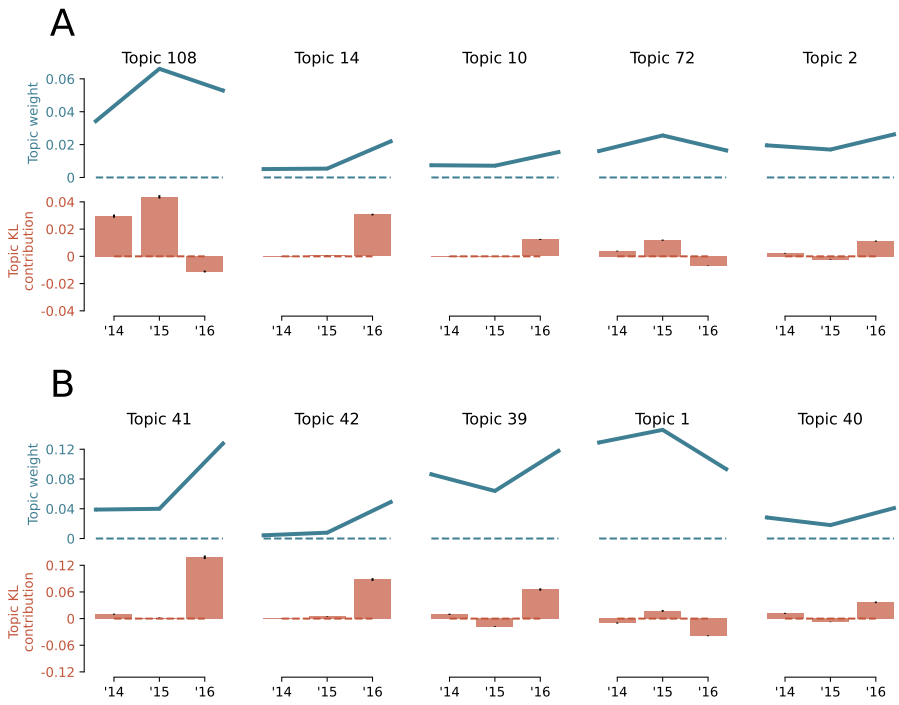


FIGURE 3.7: **Topics in housing-related decisions that change abruptly and contribute to disruption in 2016**. We show the weights of each topic (blue line) and their contribution to the Kullback-Leibler (KL) divergence (red bars) over the years preceding 2016, where there is an important disruption in the content of decisions (Fig. 3.4). We show the top-5 topics with the highest contribution to the KL divergence in the selected period for both word (A) and legislation topics (b). Topic weights correspond to topic distributions at hierarchical level 2 for word topics, and level 1 for legislation topics. While the weight shows the relative importance a topic has in a given year, the KL contribution shows the extent to which the importance of a topic is different from the year before.

case Aziz 2013[1], the European Court of Justice (ECJ) established a doctrine involving the Directive 93/13/EEC[2] on unfair terms in consumer contracts. The directive stated that, in order to protect consumers (*consumidores*, in Spanish), the validity of possible unfair terms (*cláusulas abusivas*) included in mortgage loan contracts (*contratos de préstamo hipotecario*) could be discussed during a

[1]The Aziz Judgement, by the European Court of Justice and with identifier ECLI:EU:C:2013:164, can be found at https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=ECLI%3AECLI%3AEU%3AC%3A2013%3A164_1, last accessed February 2022.

[2]Council directive 93/13/EEC, by the Council of the European Communities. See https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31993L0013, last accessed February 2022.

mortgage enforcement procedure (which, by definition, should be fast and efficient, as all its terms should already have been reviewed by a notary public when the mortgage was arranged). Since, until then, such a discussion was not formally allowed according to Spanish civil procedure law (LEC, articles 552, 557 and 561), that case forced a law reform in Spain through Act 1/2013, which was still unclear, insufficient and led to further problems and cases: since 2013, whether the amount of default interest rate (*intereses de demora*), the debt acceleration clause (*cláusulas de vencimiento anticipado*) or the unilateral liquidation of the debt by the creditor (all of them possible unfair terms) were in accordance with European law, was left to the courts to determine on a case-by-case basis during a mortgage enforcement. This process ran in parallel with the discussion about the validity of the floor clauses (*cláusulas suelo*) in mortgage loan contracts, which was not cleared up until the decision 9-5-2013 by the Spanish Supreme Court (and, at a European level, until the ECJ decision 21-12-2016).

The events of 2013 (the Aziz case with the subsequent Spanish law reform, Act 1/2013, and the doctrine of the Supreme Court on floor clauses) explain the increase in the number of decisions related to topic 108, with a peak in 2015 (see Fig. 3.7A) which is consistent with the 2/3-year delay expected in decisions ruled by courts of appeal with respect to the first instance case. Additionally, as a counterpart to the increase in the number of first instance cases related to this topic (which 2/3 years later arrived at the courts of appeal), after 2013, the number of mortgage enforcements slowed down progressively (from 38,961 in 2013 to 24,555 in 2016) due to non-judicial agreements between mortgagees and mortgagors (Garcia-Teruel and Nasarre-Aznar, 2022).

Word topic 14 Most of the words in this topic are related to a criminal offense. For instance, we find the words plaintiff and respondent (*denunciante* and *denunciado*), which appear in a criminal procedure to refer to the parties involved (instead of claimant and defendant that we would find in a civil procedure). Indeed, this topic refers to a specific type of criminal offense (minor criminal offence, *delito leve*, was introduced in Organic Act 1/2015 in substitution for misdemeanors, *faltas*; articles 13.3 and 4 and 33.4 Criminal Code): non-violent (against people) squatting (article 245.2 of the Spanish Criminal Code). This is because this topic includes references to both what is protected through this minor criminal offense, which is possession of a property (disturbance of possession, *perturbación de la posesión*; possession endangerment, *riesgo de la posesión*); and to the role that criminal law should have in this

kind of situation, namely, that criminal law should only intervene as a last resort (principle of minimal intervention, *principio de intervención mínima*; penal intervention, *intervención penal*) as far as most situations are, at least theoretically, protected through civil law using possession claims (article 250 Spanish Civil Procedure Law). However, those mechanisms were not effective in protecting owners, many of whom resorted to filing a criminal lawsuit, which was most often dismissed by judges according to the aforementioned rule of minimal intervention. This situation contributed to the enormous increase of squatters (see legal interpretation of legislation topics below) and to a reform of the Spanish Civil Procedure Law to facilitate the civil way, which did not arrive until 2018 (by Act 5/2018).

Legislation topic 41 This topic has 96% of the weight in Article 245 of the Spanish Criminal Code, whose Section 2 includes the aforementioned minor criminal offense of squatting (*usurpación*) in a usually uninhabited dwelling (see word topic 14). The sharp increase in the use of this article in housing-related decisions in 2016 is also illustrative. During the first 5 years since the start of the global financial crisis of 2007, there were no legal dispositions to stop the crisis or to palliate its consequences, while the ones since 2012 had been very feeble and non-structural (Nasarre-Aznar, 2020). This fact, coupled with social movements that supported squatting as a solution to mitigate the housing problem, led to an increase in both criminal and civil (forced dispossessions) squatting cases. According to the Spanish General Council of the Judiciary[3], convictions for squatting-related crimes went from 420 in 2007 to more than 6,000 annually between 2016 and 2018; between 2008 and 2018 convictions for squatting grew more than ten-fold.

Legislation topic 42 The most important law article in topic 42 is *Article 82 of the Spanish Organic Law of the Judicial Power (LOPJ)*, which deals with the functions of the courts of appeal in the field of criminal law. In 2015, it was modified by Organic Law 13/2015 and the word misdemeanors (*faltas*) was substituted by minor criminal offenses (*delitos leves*), in the field of criminal offenses related to housing. The second most import law article in the topic is

[3]The Spanish General Council of the Judiciary published the statistics entitled 'Criminal, civil and labor Data', available at https://www.poderjudicial.es/cgpj/es/Temas/Estadistica-Judicial/Estadistica-por-temas/Datos-penales--civiles-y-laborales/Delitos-y-condenas/Condenados--explotacion-estadistica-del-Registro-Central-de-Penados-/, last accessed February 2022

Article 876 of the Criminal Procedural Law, which deals with the process of notification of the judicial decision. This Article was also modified to substitute the same words as in *Article 82*. Since these two articles account for 80% of the weight in the topic (see topic details in Appendix B), we can say that the topic is related principally to the mentioned modification of different criminal law norms.

The aforementioned law articles address very general aspects in the scope of criminal procedural law, which hinders the task of interpreting the raise in the number of judicial decisions citing them that we observe. However, knowing that they have been modified in relation to words that appear in word topic 14, and knowing that the importance of both topics evolved in parallel (see Fig. 3.7B, 3.7G), we can say that the observed raise in the use of these articles is linked to these modifications. All in all, this is an example of the interplay between word topics and legislation topics.

3.4 Conclusion

For centuries, humans have left traces of their stories, activities and values in books, newspapers, and transcribed speeches, among others. Legal documents, such as codes, laws and, especially, judicial decisions, are particularly useful because they reflect changes in society and the evolution of the most socially sensitive issues and debates, which are the ones that end up in court. Our study presents and validates a methodology to exploit the information contained in digitized judicial decisions and to detect, quantify, and explain social disruptions from them.

One particularity of our approach is that it can be used to analyze, simultaneously and coherently, the textual content of decisions and the use of existing legislation by judges. While words shape the discourse and the arguments, citation to existing legislation summarize the mechanisms by which judges fit their ideas into the applicable framework at each point in time; thus the importance of considering both elements. Moreover, while analyzing word topics only requires linking words to legal concepts, analyzing legislation topics is more challenging because law articles can be extensive and address very general aspects. Then, using words in word topics can leverage the interpretation of law articles present in legislation topics and vice versa, as shown in our legal analysis. Remarkably, our results show that social disruption leads to abrupt changes at both levels simultaneously, that is, judges change their

discourse at the same time that they change the legal mechanisms by which they justify their decisions.

Using the same topic modeling approach to the cited legislation has the additional advantage of facilitating explanation. Indeed, law articles, like words, are semantically related to each other at different levels. In order to produce interpretable results, it is crucial to detect these relationships and reduce the dimensionality of the legislation space; that is, to go from thousands of articles to hundreds or even tens of legislation topics, as we have defined them. In other contexts, where precedent is more important than in the Spanish system (for example, in common law countries such as the United States), it may be appropriate and useful to extend the topic modeling approach to study precedent. More broadly, using topic models for content elements other than words could also be useful in different digitized documents, such as academic papers, where one could use it to coarse-grain the list of references by clustering them into topics.

Going back to the use of judicial decisions to analyze social change, we have shown that our approach provides an accurate and robust description of the shift in the content of decisions in the case of housing-related decisions in Spain. Our approach reveals that this shift is not a shallow reorganization of subtopics, but rather a deep change affecting all levels of the topic hierarchy for both words and cited law. Finally, we have shown that the topics responsible for the sharp changes we observe in the content of housing-related decisions in 2016 can be unambiguously interpreted in terms of the legal and social context at that time.

Legal documents are, together with religious documents, among the oldest written information sources that have been preserved through history. Our work shows that it is possible to extract and interpret information from such documents, and to reveal disruptive societal events. Thus, although we analyzed very recent documents, we hope that our methodology will be useful to precisely localize and interpret historical events, even in the distant past.

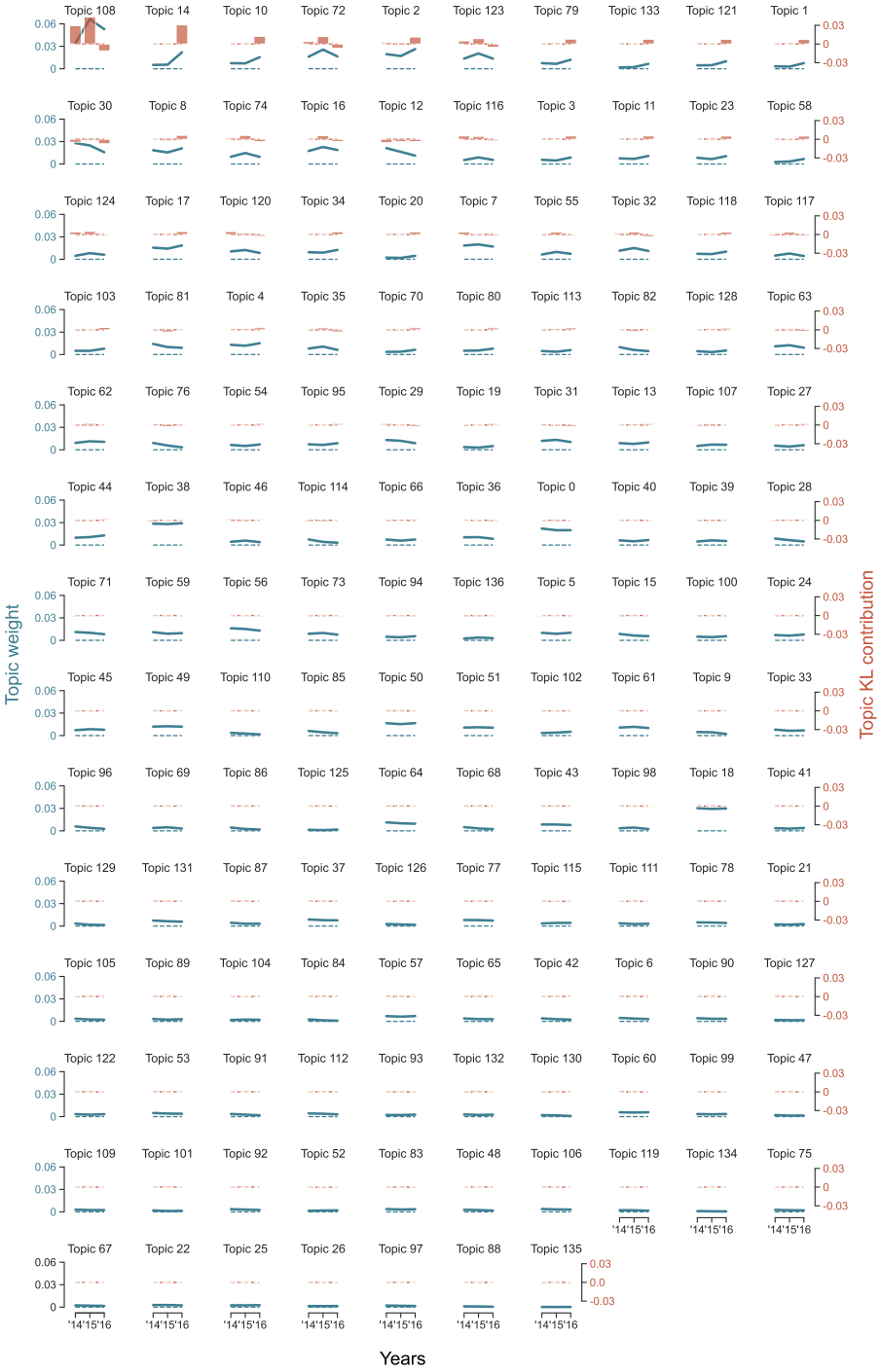


FIGURE 3.8: **Housing word topics contribution to disruption in 2016**. We show the weights (blue line) and the corresponding contribution to the Kullback-Leibler (KL) Divergence (red bars) over the years preceding 2016, where there is an important disruption in the content of decisions (see Fig. 3.7). We show all topics at hierarchical level 2.

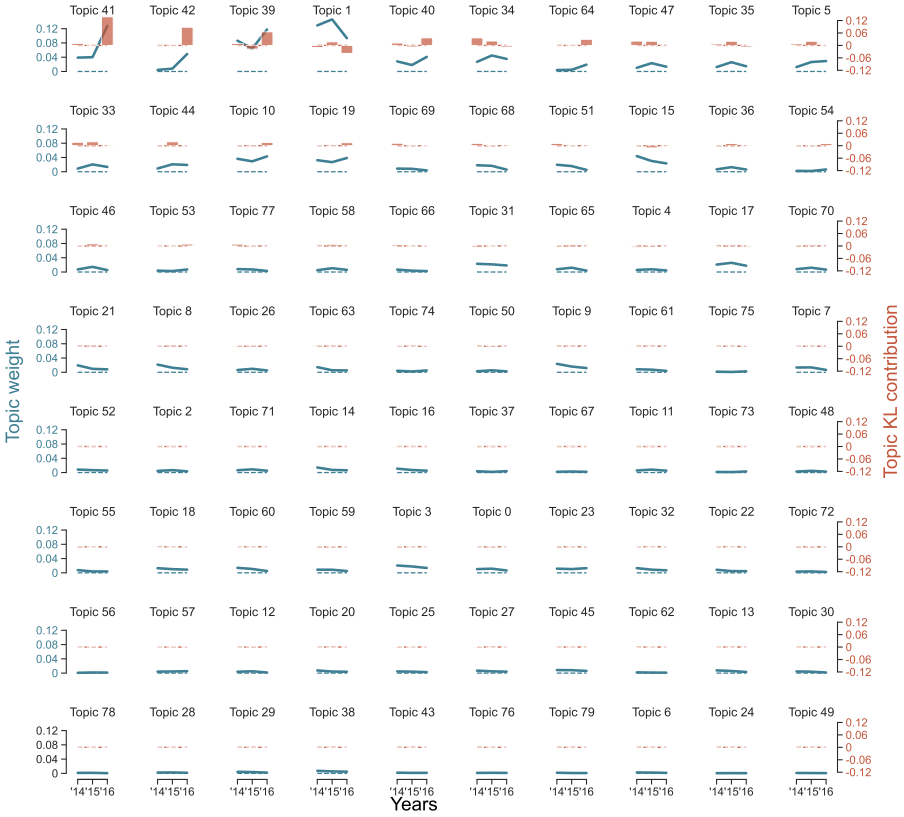


FIGURE 3.9: **Housing legislation topics contribution to disruption in 2016**. We show the weights (blue line) and the corresponding contribution to the Kullback-Leibler (KL) Divergence (red bars) over the years preceding 2016, where there is an important disruption in the content of decisions (see Fig. 3.7). We show all topics at hierarchical level 1.

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

4

Language and the use of law to predict judge gender and seniority

Social constructs and cultural stereotypes are ubiquitous and lead to unconscious bias in people's actions (Greenwald et al., 1998; Devine, 1989; Caliskan et al., 2017), even in scenarios where individuals are explicitly trained and expected to be impartial and objective, such as job interviewing (Bertrand and Mullainathan, 2004; Bagues et al., 2017), evaluation of college applications (Moss-Racusin et al., 2012), peer reviewing (Lee et al., 2013) or judicial sentencing (Danziger et al., 2011). In the scope of legal studies, many efforts have been devoted to studying the effect of judges' personal attributes on the outcome of cases, showing that gender (Asmat and Kossuth, 2021; Collins et al., 2010; Boyd et al., 2010), ethnicity (Crow and Goulette, 2022; Welch et al., 1988), age or political affiliation (Kulik et al., 2003; Cohen and Yang, 2019) can influence how cases are decided. Despite these efforts, we still lack a general theory; while some attributes such as ideology and partisanship have a clear effect on sentencing, others such as gender and race present mixed or inconclusive effects (Harris and Sen, 2019; Collins et al., 2010; Boyd et al., 2010; Eck

and Crabtree, 2020).

Justices determine the relative position of certain facts, events and actions in relation to the current applicable law. The rationale for the ultimate decision in legal cases is made explicit in the text of the judicial decisions. Thus, going beyond case outcomes by studying more subtle differences in the content of such documents can help to understand how individual and group differences among judges intrude in the legal process (Ash et al., 2022; Rice et al., 2019).

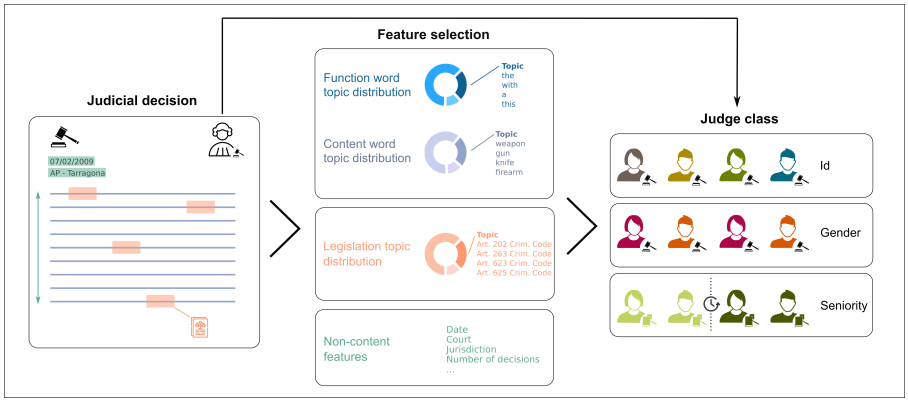


FIGURE 4.1: **Using the content of a judicial decision to predict the identity, the gender and the seniority of judges**. From the content and metadata of each decision, we take the words from the text, classifying them between function words (blue) and content words (purple) Gerlach et al., 2019, and the articles in the law cited in the text (orange). We featurize these properties by means of a topic model for each type of content Gerlach et al., 2018, so as to represent each decision as three distributions: one over function word topics, one over content word topics, and one over legislation topics. We also consider non-content features (green) by taking the date of the decision, the court, the jurisdiction and other similar features (see Methods and Data). Then, we use these features to train random forest classifiers and predict the attributes of judges: identity, gender, and seniority.

Indeed, linguistic and textual differences are predictive of the author of literary texts (Neidorf et al., 2019; Cafiero and Camps, 2019; Ainsworth and Juola, 2018), and also predictive demographic group attributes such as gender of social media content authors (Hosseini and Tammimy, 2016; Bamman et al., 2014), and age and mental state of anonymous texts (Argamon et al., 2009). Author profiling is arguably a useful task by itself – for instance, in forensics, it is a requirement for using written evidence in criminal cases (Argamon et

al., 2009; Ainsworth and Juola, 2018). However, the ability to reveal the demographic attributes of the authors from the written content they generate provides a way to unveil the inherent differences that exist between the corresponding demographic groups. In most of the previously mentioned domains, group differences are more pronounced in those aspects concerning the style of the text (Juola, 2008; Kestemont, 2014; Newman et al., 2008); however, in a few examples, group differences are more pronounced in content-related aspects such as the main ideas or the topics discussed (Koning et al., 2021; Jockers and Mimno, 2013).

When writing decisions, reporting judges tend to display a recognizable style in the form of paraphrasable content which can be expressed in formal or informal language without changing the meaning (Posner, 1995). Then, given that judges are constrained by the law and that they do not participate in the case assignation process, finding differences that go beyond style might be linked to bias in the judicial process. In this chapter, we explore whether there are measurable differences linked to the attributes of judges that translate into the content of decisions. We then investigate the extent to which these differences are just stylistic or instead substantial to the legal content. To do so, we take the three large corpora introduced in Chapter 2, that includes almost 100K judicial decisions and correspond to three legal fields in the Spanish judicial system: homicides, condominium, and housing (see section 2.2.1). We then extract features that characterize different aspects of decisions (Fig. 4.1): (i) stylistic and non-content features, such as function words (Gerlach et al., 2019; Manning et al., 2008), court ID or year of the decision; and (ii) content-related features, such as content words and references to the law. We then consider the attributes of reporting judges (their identity, gender and seniority) and measure the extent to which each of the mentioned features is predictive of these judge attributes. To do so, we use a random forest classifier that learns the values of the features that best discriminate between attribute groups. Our results show that there are strong individual differences that allow us to clearly predict the identity of the reporting judge. These differences concern stylistic and non-content features as well as content-related features. In the case of gender and seniority, while not so strong, we still find differences that go beyond writing style, allowing us to predict judge attributes more accurately than expected by chance.

4.1 Feature selection and judge attributes from judicial decisions

We use the three corpora of judicial decisions described in section 2.2.1, which include decisions from housing, homicides and condominium respectively. Most of the decisions in our data set are ruled by a group of justices that deliberate on the outcome of the case[1]. However, the reporting judge is the one in charge of writing and proposing the decision for other justices to agree on. Given the responsibility that comes with the action of writing and proposing the decision, we consider the reporting judge to be the one being analyzed. Regarding the attributes of the reporting judge, we consider the identity, the gender and the seniority (see section 2.2.1 for more details on how attributes are defined and extracted from the text), (Fig.4.1).

We quantify the content of decisions by extracting features that capture aspects of the content of decisions that range from those more linked to the legal practice and legal reasoning to those more linked to writing style. Before extracting these features, we process the text following the steps detailed in section 2.2.2. Specifically, we disambiguate specific legal-related terms and we remove numbers and non-word characters. We also *degenderize* the text by substituting all person names by '_persona_' and by removing the gender declination of certain words that mostly correlate with the gender and identity of the judge, such as *magistrado/magistrada* (masculine and feminine versions of 'justice'). We apply the *degenderization* process when using the data to predict the gender and identity of the reporting judge, but not to predict the seniority. We also find and substitute the most significant chains of words (2-grams and 3-grams), which allows us to go beyond the bag-of-words assumption and consider concepts such as 'código_civil' (civil code) or 'tribunal_supremo' (Supreme Court). For more details on the text processing steps, see section 2.2.2.

Once processed, we obtain four sets of features that characterize several aspects of judicial decisions: content-word topics, function-word topics, legislation topics and non-content features (Fig.4.1). We perform the following feature extraction processes on each corpus separately.

Content-Word topic model We filter words by using an information-theory based method to remove the most entropic words. This method is a universal,

[1] See articles 196-198, in the Spanish Organic Law of The Judiciary *Ley Orgánica del Poder Judicial*), https://www.boe.es/eli/es/lo/1985/07/01/6/con

corpus-dependent method that removes the so-called function words (also called stop words in the literature) while keeping the 'content' words, that is, more meaningful words that matter for the substantial content of documents and that improve the quality of the topics inferred afterward (see section 2.2.2 and Gerlach et al., 2019 for more detail). We also remove words appearing in less than 1% of the documents. We fix the information-content threshold to $I = 0.55$ and we take all terms below the threshold as function words and the opposite for content words.

With each document as a list of terms (significant content words and 2,3-grams), we take a topic model approach to reduce the dimensionality of the data. As in the previous chapter (3), we use the approach by Gerlach et al. Gerlach et al., 2018 to infer the topics present in the corpus and then express each document as a distribution over the topics and represent each judicial decision as a vector of weights for each topic (see section 2.2.3). Although the model is hierarchical, we select the lower level because it is the one that is more descriptive.

Function-Word topic model Because function words tend to carry stylistic signatures predictive of the attributes of written text authors (Ainsworth and Juola, 2018; Hughes et al., 2012), we also consider the stylistic content of decisions by obtaining topics of function words. Specifically, we take the words disregarded in the content-word feature extraction above, and we infer a topic model over them in an analogous way, obtaining each decision represented as a vector of the weights of each topic.

Legislation topic model Given each decision, we take the list of articles of the law cited in the text. Then, we consider an analogous approach to that of content and function words: we infer legislation topics as groups of articles in the law used similarly over the corpus of decisions, representing each judicial decision by the list of topic weights in the legislation topic distribution (see section 2.2.3).

Non-content features We also consider simple features that are not related to the specific content of judicial decisions to see how these perform in the prediction task in relation to hundreds of content-related and function-word topics. For each judicial decision, we consider: the date, the jurisdiction (civil/criminal) and the court of ruling. Besides, we also consider the number of decisions each judge has in the corpus, as an indicator of how 'prolific'

or 'experienced' a judge is in the field (Fig.4.1). To control for non-content features that have predictive ability for the same task, we benchmark against non-content features.

4.2 Judge attribute prediction from decision content-related features

Given each set of features, we evaluate the extent to which they are informative in predicting the discretized attribute (identity, gender, seniority) of the judge. To do so, we train a random forest algorithm, a supervised classifier that uses an ensemble of decision trees and learns how to classify the data from the features. The algorithm is well suited for classification problems with high-dimensional data, and it has been widely applied in a variety of domains (Breiman, 2001; Boulesteix et al., 2012). Then, we validate the trained classifier using a K-fold cross-validation, that is, dividing our data set into K portions, training the classifier on $K-1$ portions while testing in the resting one, and repeating for all K combinations of train and test sets. The number of folds chosen in each case tries to balance performance and computational cost. The decisions are randomly assigned to each portion, while keeping the proportion of classes equal to the global one. In the same sense, the predictions are calibrated to ensure that the proportion of predicted classes is equal to the proportion of the data.

4.2.1 Judge identity is highly predictable from language and use of legislation

We start by analyzing the predictability of the identity of a judge from the content of their decisions. Considering a 10-fold cross-validation, we take a subset of decisions where each judge has at least 10 decisions to ensure the presence of each judge in all 10 portions at least once. In the case of condominium, where we have a much larger corpus, we consider a threshold of 70 decisions to ensure the computational feasibility of the random forest classifier. Figure 4.2A-C shows the accuracy in the prediction of the identity of the reporting judge from features that capture the legal aspects of decisions, namely, content word topics and legislation topics.

From content-word topics, and considering first the homicides corpus, we can predict the exact identity of the judge in 63% of the decisions when using the set of decisions from judges with at least 10 decisions in the corpus

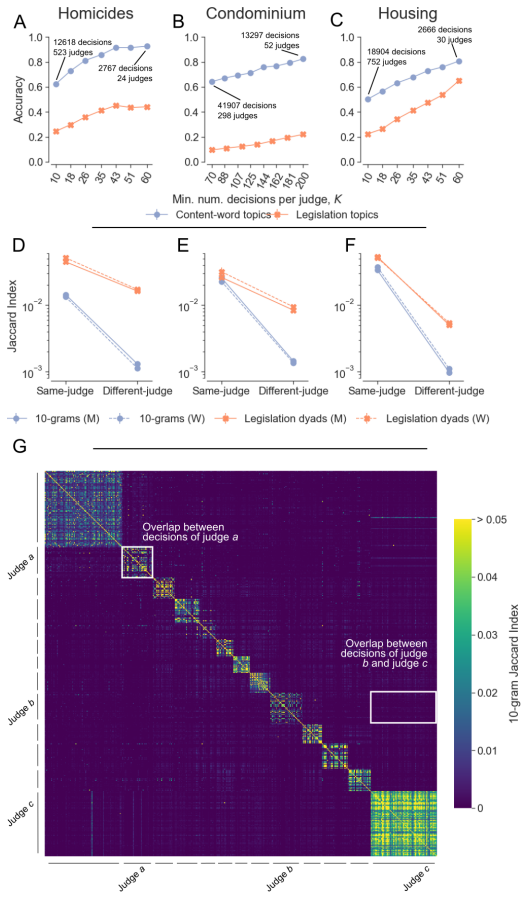


FIGURE 4.2: **Differences in language use and law citation are highly predictive of the reporting judge.** (A-C) Accuracy in the prediction of the identity of the reporting judge for the homicides (A), condominium (B), and housing (C) corpora. We use content-word topics and legislation topics as predictive features. For each corpus, we consider different subsets of decisions, each corresponding to decisions from judges with a minimum number of decisions per judge, K. We report the accuracy of the predictions, that is, the fraction of times that the judge identity is predicted correctly. Results are averages over a 10-fold cross-validation. (D-F) Degree of content overlap between pairs of decisions for the homicides (D), condominium (E), and housing (F) corpora. For each decision, we consider: (i) the list of consecutive 10 words (10-grams); and (ii) the list of legislation dyads. We compute the average Jaccard index between same-judge decision pairs and different-judge pairs of decisions. We also differentiate between male and female reporting judges (M/W). Standard error bars computed over folds are smaller than symbols in all plots (A-F). (G) 10-grams Jaccard index between decisions from a selection of 13 judges in the homicides corpus.

(12,618 decisions from 523 judges). When we restrict the analysis to judges with at least 60 decisions (2,767 decisions from 24 judges), the accuracy goes up to 93%. The results are similar for the other two corpora: in the condominium corpus, we obtain 64% accuracy for judges with at least ten decisions (41,907 decisions from 298 judges), and 82% accuracy for judges with at least 60 decisions (13,297 decisions from 52 judges); in the housing corpus, we obtain 50% accuracy (15,331 decisions from 664 judges) and 81% accuracy (1,564 decisions from 17 judges), respectively. Legislation topics are also very predictive of judge identity, although less so than content-word topics. Using a similar selection of decisions (only disregarding a small fraction of decisions with no legislation cited) for each corpus, we achieve accuracies in the range from 25% to 44% in homicides; 10% to 22%, in condominium; and 22% to 65% in housing. In both cases, results are much higher than what would be expected by chance (using a calibrated naive guesser[2], we obtain an accuracy of 0.4% to 4% in homicides, 0.4% to 2% in condominium and 0.2% to 6% in housing).

Judge identity prediction is linked to content reuse

These results suggest that judges must have strong individual signatures that affect both the use of content words and the references to the law, something that makes them very recognizable from a classification point of view. To take a closer look at these individual signatures, we analyze the degree of overlap that exists between pairs of decisions. We hypothesize that this recognizable signature should arise from a higher overlap between pairs of decisions from the same judge (same-judge pairs) compared to that between pairs of decisions from different judges (different-judge pairs). Again, to compare decisions, we use both words and cited legislation. Specifically we consider: (i) chains of consecutive 10 words (10-grams), disregarding chains that include punctuation marks (except before the first word or after the last one) and those only appearing in just one decision; and (ii) combinations of pairs of cited law articles (legislation dyads), considering the list of references to articles in the law and taking all possible pairs.

[2]We compare our results for the prediction tasks of the identity, gender, and seniority of the judge with a null model characterized by a calibrated naive guesser, which is equivalent to a random assignation of judge attribute labels in the test set while preserving the ratios of each class, and the subsequent performance evaluation in terms of the accuracy (judge identity) or AUROC (judge gender and seniority).

Being W_d and L_d the set of 10-grams and the set of legislation dyads corresponding to decision d, respectively, we measure the normalized intersection between two decisions d and r using the Jaccard index:

$$J_{dr}^{W} = \frac{|W_d \cap W_r|}{|W_d \cup W_r|}, \tag{4.1}$$

$$J_{dr}^{L} = \frac{|L_d \cap L_r|}{|L_d \cup L_r|}. \tag{4.2}$$

We then compute the degree of overlap corresponding to each judge. To that end, we consider the set of decisions D_i written by judge i. Then, to estimate the reuse of content from own decisions, we first compute the degree of overlap between decisions within this set.

$$J_i^{X_{\text{self}}} = \frac{2}{|D_i|\,(|D_i|-1)} \sum_{(d,r)\in D_i, d\neq r} J_{dr}^{X} \qquad \text{with} \quad X = \{L, W\}. \tag{4.3}$$

Second, we consider the overlap between decisions D_i written by judge i and the decisions written by other judges, $D_{\neq i} = \{\bigcup_{j\neq i} D_j\}$,

$$J_i^{X_{\text{other}}} = \frac{1}{|D_i|\,|D_{\neq i}|} \sum_{d\in D_i; r\in D_{\neq i}} J_{dr}^{X} \quad \text{with} \quad X = \{L, W\}\,. \tag{4.4}$$

Globally, we average these quantities over all judges:

$$\langle J^{X_{\text{self}}} \rangle = \frac{1}{N_{\text{judges}}} \sum_i J_i^{X_{\text{self}}} \quad , \qquad \langle J^{X_{\text{other}}} \rangle = \frac{1}{N_{\text{judges}}} \sum_i J_i^{X_{\text{other}}} \quad . \tag{4.5}$$

In other words, these previous expressions (4.3 and 4.4) tell us the extent to which we find content (in the form of either chains of words or pairs of cited legislation) appearing in more than one decision. For instance, a value of $J_l^{W_{\text{self}}} = 0.05$ would mean that, if we took randomly a pair of decisions from a judge l and joined their text, we would expect a 5% of the 10-grams appearing in both decisions.

Finally, we measure the individual tendency of each judge reuse more their own content than from other as the difference between these quantities as:

$$\Delta J_i^{X} = J_i^{X_{\text{self}}} - J_i^{X_{\text{other}}}\,. \tag{4.6}$$

The results from a selection of judges in the corpus of homicides (see Fig. 4.2G for 10-grams overlap) show that the degree of overlap between same-judge pairs is much higher than that between different-judge pairs, which results in a distinct diagonal pattern in the matrix of overlaps. Extending the analysis to the other corpora and to legislation dyads, we obtain the same result (Fig. 4.2D-F). For 10-grams, we observe a 15-fold increase in the degree of overlap between same-judge pairs and different-judge pairs in homicides. For condominium and housing, the increase in overlap is 20 and 40-fold, respectively. In the case of the cited legislation dyads, the increase in overlap is more modest but still sizable: 2-fold in homicides, 3-fold in condominium, and 6-fold in housing. In terms of the gender of the reporting judge, there is no appreciable difference between men and women.

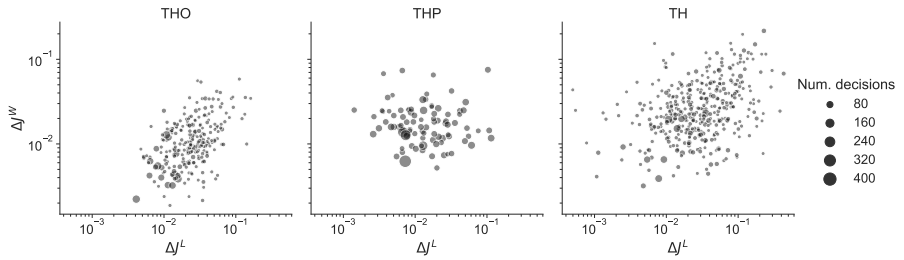


FIGURE 4.3: **Correlation between word reuse and legislation reuse.** Given the individual tendency of each judge reuse more their own content than from others, ΔJ, we plot the corresponding measure for the reuse of words, ΔJ^W and for the reuse of cited legislation, ΔJ^L. See Methods in Main text. Spearmann correlation: Homicides: 0.426, p-value:$1.76x10^{-12}$; Condominium: -0.133, p-value:0.20; Housing: 0.343, p-value:$3.59x10^{-12}$.

Among judges, the tendency to reuse more words from their own decisions than from others' (ΔJ_i^W in expression (4.6)) seems to be positively correlated with the tendency to also reuse cited legislation from their own decisions more than from others' (ΔJ_i^L, see Fig. 4.3; not significant in the condominium corpus). This suggests that these two observations are two sides of the same coin of content reuse.

4.2.2 Judge gender can be predicted from content-related features

Our results clearly show that there are individual traces in each decision that make it possible to guess the identity of the author of each decision. However, this finding does not answer the question of whether there are also more generic group signatures in the content of decisions that allow to identify attributes of judges such as gender or seniority. In what follows, and to prevent the classifier from learning the gender and seniority of judges by first learning their identity, we aggregate all the decisions of each judge, computing the average distributions over word and legislation topics and the average over non-content features. In this way, each judge is represented by a single *average decision*. In all forthcoming cross-validation experiments, we split judges (and their average decisions) into training and validation sets. Therefore, we predict the gender and seniority of each judge from a training set that only includes the decisions of other judges, but not their own, so that identity cannot possibly be learned.

Similarly to the case of predicting the identity of a judge, we evaluate how the differences in the gender of the reporting judge translate into differences in the content of decisions. We find that both content-word topics and legislation topics can be used to predict the gender of the judge better than expected by chance (Fig. 4.4A-C). In the case of content-word topics, the area under the receiver operating curve (AUROC) ranges from 0.59 in housing to 0.69 in homicides (0.62 in condominium). In the case of legislation topics, the AUROC ranges from 0.54 in housing to 0.61 in homicides (0.55 in condominium; see Fig. 4.4A). The results for the F1 metric also show both features performing better than chance (Fig. 4.4B,C). Additionally, we find that content word topics are more predictive than legislation topics for the three corpora: (AUROC is 12% higher in homicides, 17% higher in condominium and 11% higher in housing; Fig. 4.4A. Similar results hold for F1 metrics; Fig. 4.4B, C). These results show that there are inherent differences between male and female judges that permeate into measurable differences in the content of decisions they write, differences that even allow us to predict the gender of the judge better than expected by chance.

To benchmark the predictive power of content features (content words and legislation), we compare them to the predictive power of non-content features such as the date of the decision, the ruling court, and the number of decisions each judge has in the corpus. While there are no significant gender

differences in the number of decisions per judge (see Fig. 2.4), the differences regarding the ratio between decisions written by men and women vary considerably, both over years and across courts. For example, in 2001, only 3% of the decisions in the homicides corpus were written by women, while they amounted to 34% of decisions in 2018 (see Fig. 2.5 for more details). Similarly, whereas just 1% of the homicides decisions ruled by the Supreme Court were written by women, the fraction goes up to 50% in Madrid's Provincial Audience (see Figs. 2.6, 2.7, 2.8). Given these marked differences, it is clear that this information should help considerably to predict the gender of the judge better than expected from chance; we confirm this expectation (Fig. 4.4A-C). The performance comparison between content and non-content features shows that word topics perform better in the condominium (13% better) and the homicides corpora (19% better), and similarly in the housing corpus. In the case of legislation topics, the performance is equivalent in homicides and condominium, and 9% lower in housing (see Fig. 4.4D-F). Therefore, even though non-content features are intuitively quite predictive because of the large gender disparities in time and geography, content features (especially content words) are often even more predictive or, at least, similarly predictive (with only the exception of legislation features in the housing corpus).

4.2.3 Judge seniority can be predicted from content-related features

Along the same lines of gender prediction, we explore how the differences in the content of judicial decisions are predictive of the seniority of judges, that is, the length of their careers as measured by the number of years of service. To maintain the structure of the prediction task with respect to gender, we split judges in two groups according to their seniority: early-career judges and senior judges.

Results in Fig. 4.4D-F show that content word topics can predict the seniority of the judge more than expected by chance (AUROC scores: 0.58 in housing, 0.61 in condominium, and 0.62 in homicides). In the case of legislation topics, results are similar except for the case of housing, where results are not distinguishable from chance (AUROC scores of 0.60 in condominium and homicides and 0.48 in housing).

Similarly to the case of gender prediction, we compare the performance of content topics with non-content features. In this case, non-content features include the court and the number of decisions by each judge during the time

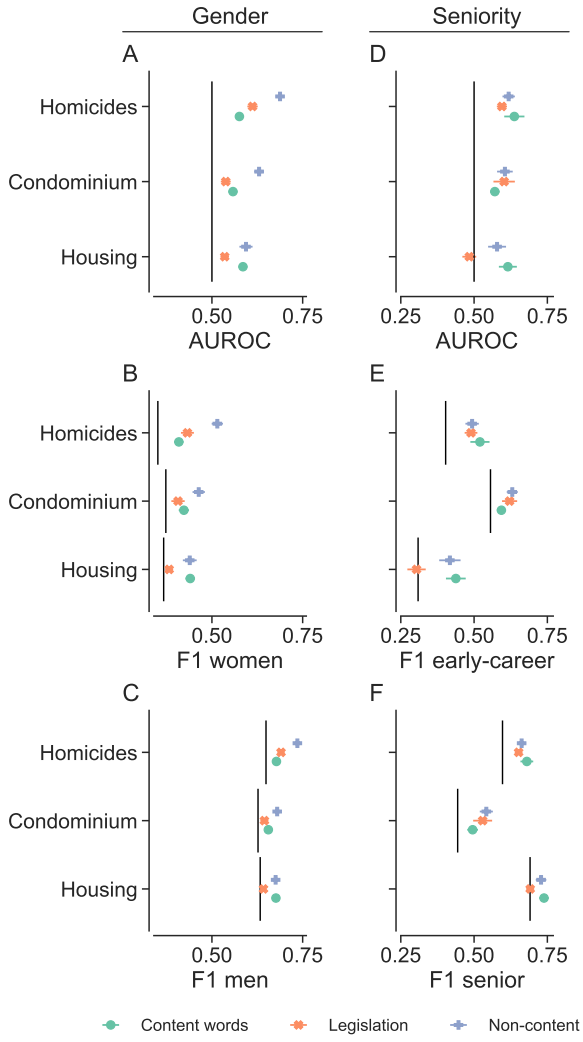


FIGURE 4.4: **Judge gender and seniority prediction.** We predict the gender and the seniority of the reporting judge using different features taken from the content of their judicial decisions: topics of content words, topics of legislation, and non-content features (see Methods). We train a random forest classifier and we evaluate the prediction using three different metrics: the area under the receiver operating curve (AUROC - **A, D**), the F1 score for the 'women'/'early career' class (**B, E**) and for the 'men'/'senior' class (**C, F**). We show the score of each metric compared to a calibrated naive guesser (black line, see Methods). Each point corresponds to the average of a k-fold cross validation (k =20 for gender, k =15 for seniority) and error bars represent the standard error of the mean. Where not visible, error bars are smaller than symbols.

window we consider. The ratio of senior to early-career judges varies considerably across courts, going from 31% of early-career judges in the Supreme Court to 85% in Barcelona Provincial Audience for homicides (see Figs. 2.11, 2.9 and 2.10). The performance for both content-word and legislation topics being statistically indistinguishable from that of non-content features in all three corpora (except for legislation in housing) gives an idea, again, of the extent to which seniority differences affect the legal content of decisions.

4.2.4 Function words are as predictive of seniority and more predictive of gender than content word

Up to this point, we have analyzed how features related to legal reasoning (content words) and legal practice (cited legislation) are predictive of the gender and seniority of reporting judges. Next, we analyze the predictive power of features that characterize the stylistic aspects of the text of decisions Langford et al., 2020; Kestemont, 2014, considering topics of function words.

Our analysis shows that function words are predictive of the gender of the judge, with AUROC scores of 0.72 in homicides, 0.73 in condominium and 0.63 in housing (Fig. 4.5). Similarly, results show that function-word topics are also predictive of judge seniority above what is expected by chance in all three corpora, with AUROC values of 0.63 in homicides, 0.65 in condominium and 0.56 in housing.

To benchmark these results against those obtained using content words, we calculate the log-ratio between the predictive accuracy of function words and that of content words (Fig. 4.5B and D). Positive log-ratios indicate that function words are more predictive than content words, and vice versa. We find that function words are as predictive of judge seniority as content words. However, function words are significantly and consistently more predictive of gender than content words, with log-ratio values of 0.04, 0.16 and 0.07 for the homicides, condominium, and housing corpora, respectively. Results are averaged over a k-fold cross-validation ($k = 20$ for gender, $k = 15$ for seniority).

Considering alternative thresholds for the information-content value to separate content words from function words, we find similar results. For gender prediction, using function words is always much more predictive in the condominium and housing corpora, while the improvement is more modest in

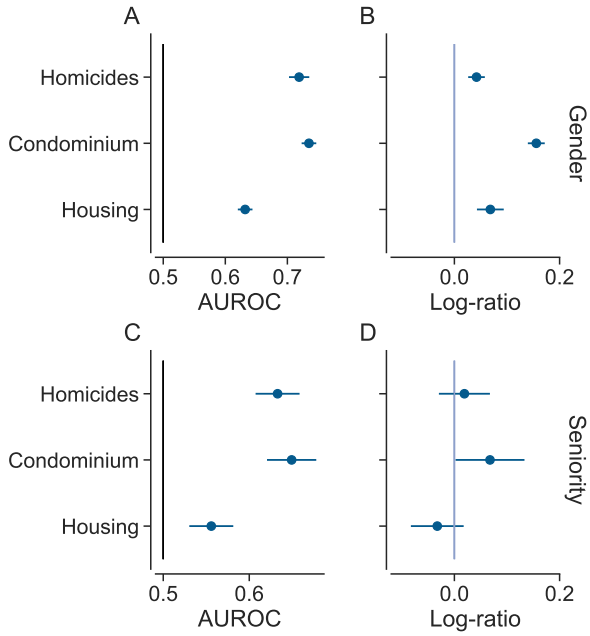


FIGURE 4.5: **Function words are as predictive of seniority and more predictive of gender than content words.** (A,B) Using a random forest classifier, we predict the class of the judge (gender or seniority) using function word topics. In particular, we show the area under the receiver operating curve (AUROC) for the prediction of the gender (A) and seniority (C). (B,D) To compare the predictive power of function words to that of content words (as reported in Fig. 4.4), we show the log-ratio between the AUROC for the prediction with function word topics and that of content word topics. Each point and error bar correspond to the average and standard error of a k-fold cross-validations (k =20 for gender and k =15 for seniority). Where not visible, error bars are smaller than symbols. See Fig. S9-10 for the corresponding results for F1 scores.

the homicides corpus. In the case of seniority, using function words or content words has the same predictive power (see Figs. 4.6 and 4.7).

4.2.5 Gender and seniority differences are attributed to complex combinations of features

The results we presented in previous sections show that we can predict, better than would be expected by chance, the gender and seniority of judges from the topics that quantify the judicial decisions they write. However, we achieved these predictions by using topic models that involve a considerable

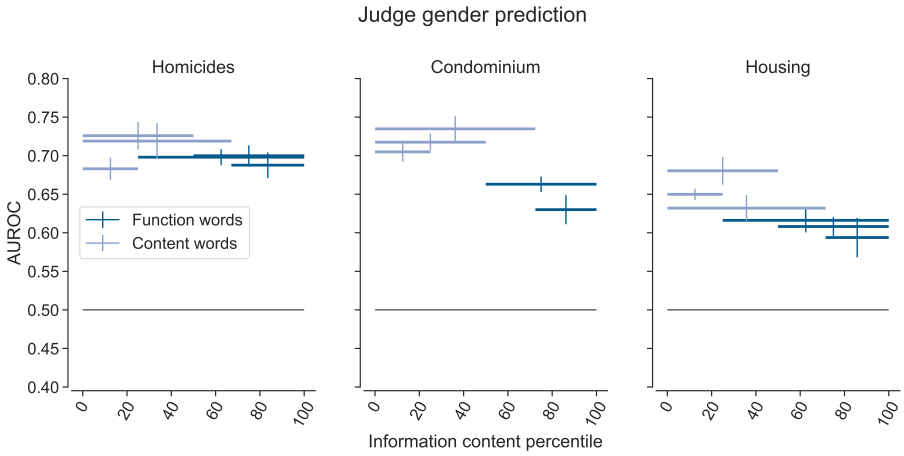


FIGURE 4.6: **Judge gender prediction performance for content and function words for different information-content thresholds.** We classify the terms appearing in each corpus decisions between content and function words according to the information-theoretic criteria detailed in Gerlach et al., 2019. We use each content/function words selection to predict the gender of judges as described in main text. In the figure, the span of each horizontal line represents the span of the information-content values for the terms used in the prediction, in percentiles over the information-content distribution of the corpus. The results for the area under the receiver operating curve are averaged over a 10-fold cross-validation.

number of topics (more than 10^3 for content word topics, more than 10^2 for legislation word topics). Therefore, we wonder if there exists a considerably smaller subset of these topics that could achieve a similar prediction performance and thus facilitate the interpretation of the results. For this reason, we took several different-sized subsets of topics and we evaluated their predictive power.

Taking the set of judges to predict $j \in [1, J]$, and the average topic distribution for the decisions of each one of them, we compute the correlation between the weights of each topic and the attributes of each judge, measured as the mutual information between these two functions. Specifically, and for the case of the prediction of judge gender:

$$I(G, T_k^X) = \sum_{i \in J} \sum_{j \in J} P(G, T_k^X | i, j) \log \left(\frac{P(G, T_k^X | i, j)}{P(G|i) P(T_k^X | j)} \right) , \tag{4.7}$$

where $P(T_k^X | j)$ is the weight of a specific topic k in the average judge topic

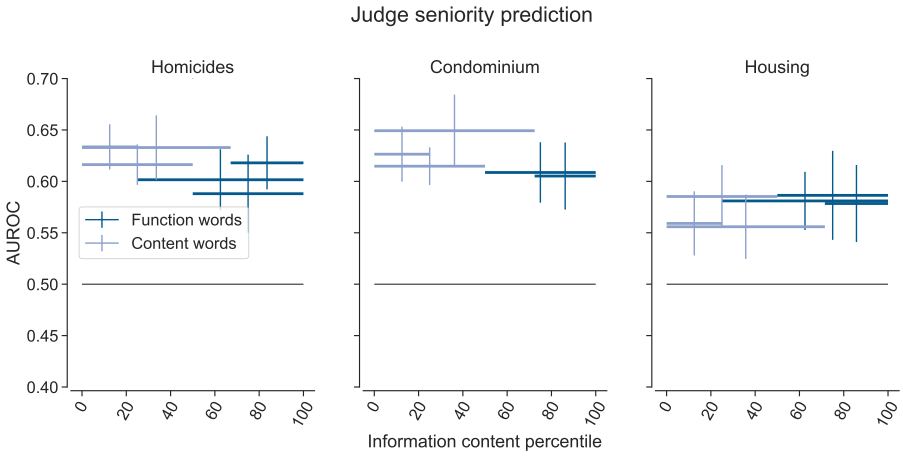


FIGURE 4.7: **Judge seniority prediction performance for content and function words for different information-content thresholds.** We classify the terms appearing in each corpus decisions between content and function words according to the information-theoretic criteria detailed in Gerlach et al., 2019. We use each content/function words selection to predict the seniority of judges as described in main text. In the figure, the span of each horizontal line represents the span of the information-content values for the terms used in the prediction, in percentiles over the information-content distribution of the corpus. The results for the area under the receiver operating curve are averaged over a 10-fold cross-validation.

distribution, $P(G|j)$ is the function that sets the gender of the judge, and $P(G, T_k^X|i,j)$ is the joint probability mass function of both functions. $X \in \{L, W\}$ represents either using content-word topics or legislation topics. Similarly, we can compute the analogous mutual information for the seniority of the judge:

$$I(S, T_k^X) = \sum_{i \in J} \sum_{j \in J} P(S, T_k^X|i,j) \log \left(\frac{P(S, T_k^X|i,j)}{P(S|i)P(T_k^X|j)} \right) , \tag{4.8}$$

where $P(S|j)$ is the function that sets the seniority of the judge. Once computed the correlation for each topic, we descendantly order them, and then we select subsets, taking the first N.

In Fig. 4.8 we show the results regarding the gender prediction performance. When using content-word topics, the performance falls systematically when reducing the number of topics, and when taking 10 topics it falls from an AUROC score of 0.69 to 0.55 in homicides, from 0.64 to 0.55 in condominium, and

from 0.58 to 0.53 in housing. When using legislation topics, the results fluctuate more than those of gender prediction. In homicides, the performance falls from 0.60 to 0.55% when reducing to 10 topics, but the performance is 0.60 for a set of 8 topics and 0.53 when reducing to 4 topics. In Condominium and Housing, the performance level fluctuates around the score resulting from considering all topics (0.55 in condominium, 0.53 in housing) and eventually falls when using 4 topics ((0.53 in condominium, 0.50 in housing).

In Fig. 4.9 we show the results regarding the seniority prediction performance. When using content-word topics, the results different sets of topics fluctuate considerably. In homicides, where the performance using all topics is an AUROC score of 0.58, a selection of 629 topics results in a score of 0.68 whereas a selection of 40 topics gives a score of 0.54. In the case of condominium and housing we find a similar behavior. In all three cases the performance drops below 10 topics (AUROC score of 0.55 using 7 topics in homicides, 0.51 using 6 topics in condominium and 0.52 using 7 topics in housing). When using legislation topics, the performance in homicides fluctuates around the score corresponding to using all topics (0.58) and eventually drops to 0.56 using 4 topics. In the case of condominium, the performance falls from AUROC score of 0.6 using all topics to 0.57 using 4 topics. In the case of housing the score using all topics falls bellow the expected by chance (0.49) and exploring the performance over the different sets of topics produces results that fluctuate from 0.61 using 30 topics to 0.44 using 16 topics.

These results show that we are not able to systematically maintain the same prediction performance when reducing the number of topics considerably. In fact, either the performance drops systematically or it fluctuates showing inconclusive results. All in all, these results imply that we are not able to explain the gender and seniority differences in terms of a set of a few topics. Contrarily, these differences are the result of a complex and intrincated combination of many topics.

4.3 Conclusion

Our results show that there are inherent differences in the way judges write decisions, which make them recognizable not only at the individual level but also when grouping them by gender or seniority. In our analysis, we use a range of features that capture both the style and legal reasoning of decisions,

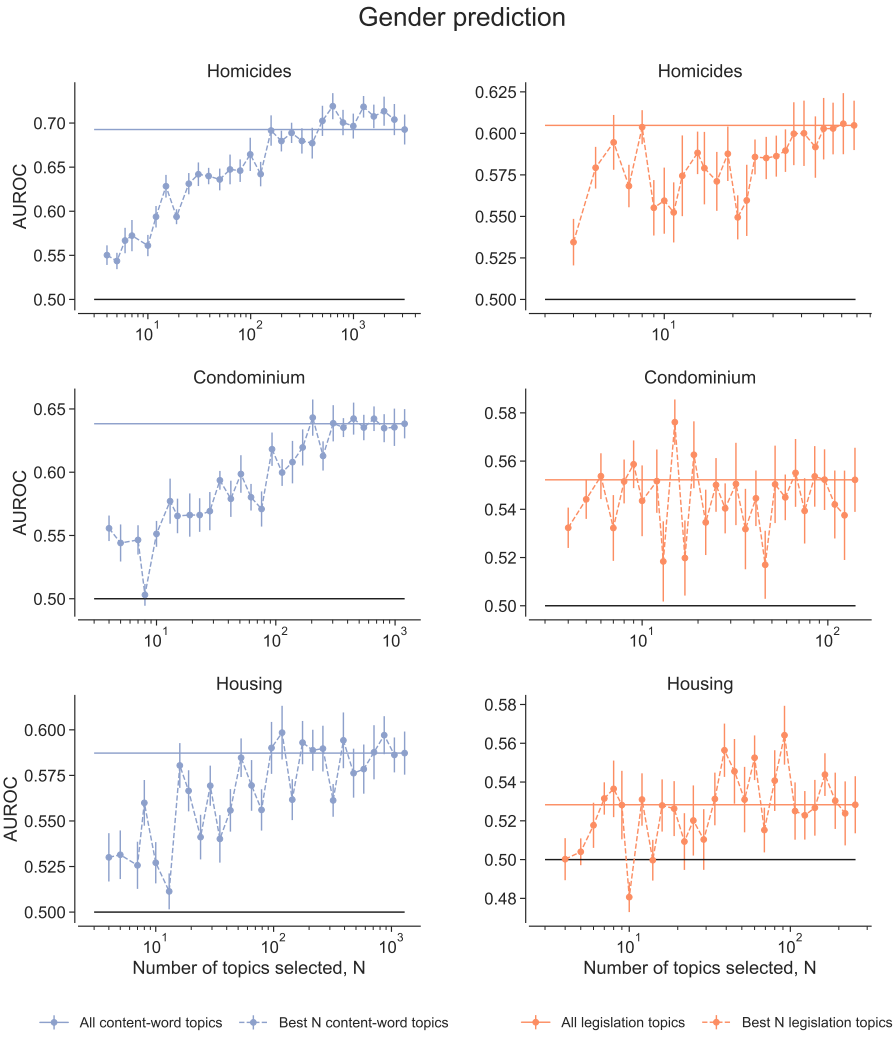


FIGURE 4.8: **Performance dependence on feature selection in gender prediction.** Using content-word topics and legislation topics to predict the judge gender, we select the N topics with the highest values of mutual information with the gender of the judges (see equation 4.7) and we evaluate the prediction performance for each feature selection in terms of the area under the receiver operating curve, AUROC.

which allows us to better understand the nature of the differences between judges.

At the individual level, our results show that one of the primary causes for the appearance of these differences is that judges reuse more content from their past decisions than content from other-judge decisions. Reasonably, this can

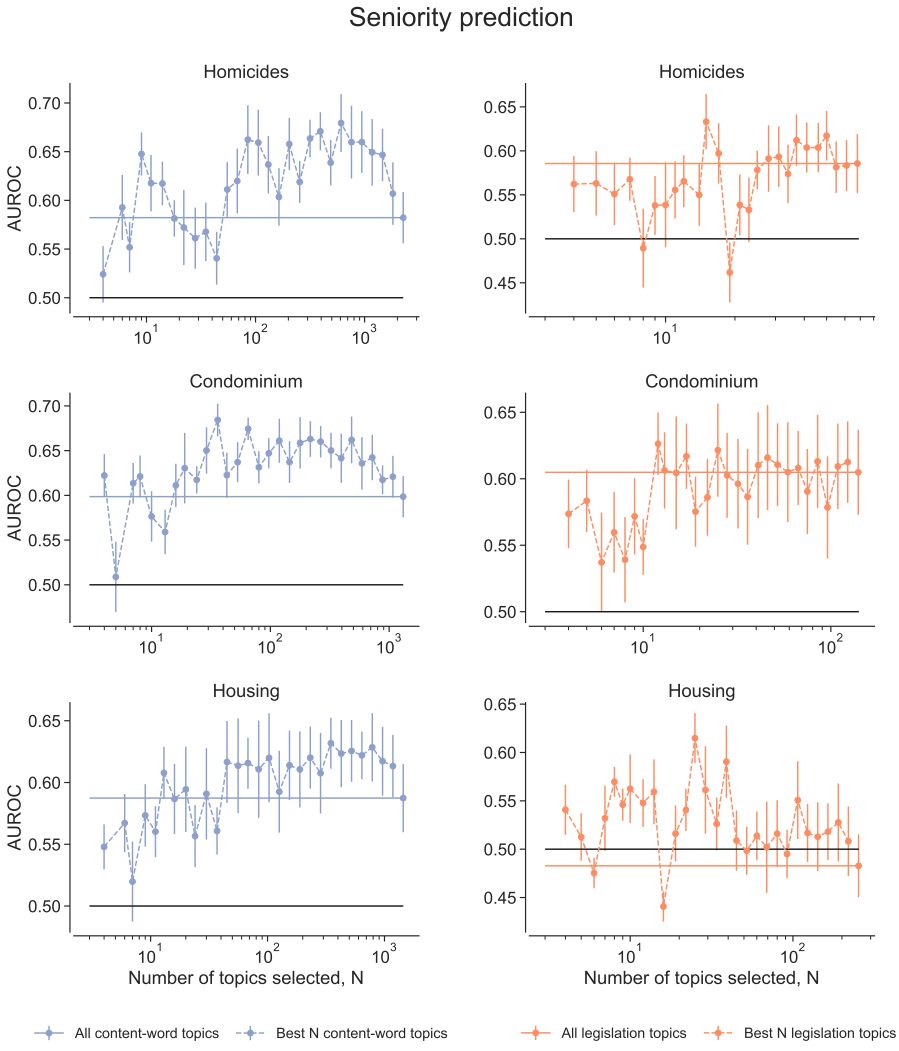


FIGURE 4.9: **Performance dependence on feature selection in seniority prediction.** Using content-word topics and legislation topics to predict the judge seniority, we select the N topics with the highest values of mutual information with the seniority of the judges (see equation 4.8) and we evaluate the prediction performance for each feature selection in terms of the area under the receiver operating curve, AUROC.

be expected given the fact that judges have their own way of saying things, using a certain tone, expressions and a given level of technical language (Posner, 1995). Moreover, when facing a new case, judges might find it easier to remember similarities with past cases of their own rather than spending time looking for similarities in the vast archive of available decisions. One can

then expect that the reuse of content translates into individual idiosyncrasies of judges' writing styles, which we are able to reveal by using function words to predict the identity of the judge. However, by only considering the content that underlies the legal reasoning and the framing of the case in the applicable law, that is, considering content words and cited legislation, we still predict very well the identity of the judge, which reveals that content reuse affects the wording of arguments and the choice of supporting legislation as well.

The reuse of content from own documents is common in other domains as well. For instance, in science, scholars tend to reuse text from papers they have authored much more than text from the papers of others (Citron and Ginsparg, 2015). This is to be expected, given that scientists have few constraints in the choice of the subject and the methodology of research, and they often draw upon their previous results to move forward. However, in our case of study, the situation is rather different: cases are randomly assigned, so that judges are not free to choose the *subject* of the cases they must decide on.

The individual traits we observe could also come from another source. Normally, courts organize themselves into different sections and chambers, to which judges are assigned. According to the Spanish Organic Law of the Judicial Power (LOPJ, art. 152.2), these courts can decide how to assign cases among sections and chambers. This assignation is based on subject criteria; judges are assigned cases depending on their domain of expertise. Thus, it could be possible to find judges whose decisions can be differentiated from those of others by the legal subject. However, we are considering three corpora of cases in very specific fields, which typically would be assigned to experts in their respective courts, and therefore, these thematic differences cannot explain our results.

Beyond individual differences, our analysis in terms of features of the set of decisions of each judge reveals differences that are predictive of both gender and seniority of judges. We observe these differences for both content and non-content related features. However, despite our reduction of the dimensionality of the description of decisions from tens of thousands of words to hundreds of features, we cannot pinpoint the specific sources for these differences – indeed, we find that the differences we observe are not attributable to a few features of the decisions but to complex combinations of them. Finding the sources for these differences is thus not trivial, but poses a question that should be investigated in depth. Actually, because these differences cannot

be attributed neither to individual differences nor to case assignation criteria, understanding how these observed differences translate into differences in how judges apply the law is a fundamental question that needs to be answered. Further efforts in this direction could enable an intervention in the case allocation policies in the courts, ultimately contributing to the transparency and well-functioning of the judiciary.

5

Atypical and pioneering legislation citations can predict content reuse

Legal systems, through their adherence to legal principles and precedents, aim at ensuring a consistent and coherent application of the law that guarantees public trust, transparency, fairness, and equity in the administration of justice. This coherence and consistency can be explicitly measured by the predictability of the use of legislation and justices votes to decisions throughout cases (Mones et al., 2021; Guimerà and Sales-Pardo, 2011; Medvedeva et al., 2020). At the same time, the legal system is forced to gradually change, subject to a mixture of societal, cultural, and political factors (Sheshadri and Singh, 2019; Rutherford et al., 2018; Klingenstein et al., 2014; Katz et al., 2020; Rockmore et al., 2018). These changes, which arise in the context of the judicial system functioning as a socially interacting system, can be globally understood from the small and individual contributions made by novel and uncommon behaviors.

In this chapter, we study how to quantitatively define and detect innovative practices in the judicial system through the study of judicial decisions. To do

so, we adapt approaches from the area of science of science which have been used to study innovation using different definitions, such as atipicality (Uzzi et al., 2013), disruption (Wu et al., 2019), interdisciplinarity (Fontana, 2020) and high-stakes research (Rzhetsky et al., 2015). Since innovation has to be measured in comparison to already existing research, these studies discretize, in a network representation, the space of 'all possible research' in terms of key features of the research papers studied, such as molecules (Rzhetsky et al., 2015), genes (Stoeger et al., 2018), concepts (Iacopini et al., 2018; Krenn and Zeilinger, 2020), 'memes' (Kuhn et al., 2014), and references to previous works (Uzzi et al., 2013). Analogously, here we represent the space of 'all possible uses of legislation' by considering a network representation of the citations to law articles appearing in each judicial decision within a corpus. Then, we define novelty in terms of the already existing pattern of uses of the law in the network.

We evaluate innovative practices from different perspectives. First, adapting the definition by Uzzi et al., 2013, we measure atypical combinations of citations to legislation. Second, given that the previous definition lacks a time perspective, we introduce two new definitions based on the identification of pioneering legislation strategies: first, we group together decisions that cite legislation in a similar way by fitting a Stochastic Block Model (Peixoto, 2019) in the bipartite network of decisions and law articles and, second, we assign high values of novelty to earlier decisions in each group. In science, innovative practices have been linked to impact, and to the diffusion and inheritance of these practices (Uzzi et al., 2013; Iacopini et al., 2018; Kuhn et al., 2014). To assess whether a similar phenomenon occurs in a legal context, we evaluate the degree of adoption of the content of novelty decisions in terms of the reuse of content made by future decisions. Our results show that there is some degree of correlation between the innovative traits of a decision and the impact it will have on the future, and that this correlation is predictive of the impact of a decision given its novelty values.

5.1 Novelty in judicial decisions based on cited legislation

In what follows, we use decisions corresponding to three corpora encompassing Housing, Condominiums and Homicides decisions (see 2.2.1). Besides the main text, we have access to the extracted list of citations to legislation

and, in particular, citations to articles in the law, which we will use to define novelty practices.

5.1.1 Novelty as atypical combinations of cited legislation

We introduce a definition of novelty of a judicial decision based on the cited legislation of the decision. Our approach follows the work by Uzzi et al., 2013, where the authors quantified the novelty of scientific papers depending on the pair-wise combinations of journals found in their references. Specifically, they measured the extent to which journal pairs were commonly or uncommonly cited in a given publication year. To do so, they took all scientific papers in a given year and measured the number of times a given journal pair appeared cited. Then, they compared this frequency to the expectation of a configuration null model over the bipartite network formed by papers on one side and cited papers on the other, with links being citations from the first group to the second. A configuration null model is a random version of the network that preserves some structural characteristics (Bollobás, 2001). In this case, they shuffled citation links while preserving the total number of citations made and received by each paper, as well as the years of publication of the cited papers. In this way, the structure of the network is preserved: for instance, the most cited article is still the most cited one in the random network, or the number of citations of papers from a given year is still preserved in the random network, etc. Finally, they quantified how uncommon citing a pair of journals (i,j) in a given year y was by computing the z-score between the observed frequency in the real network and the frequency in the configuration null model:

$$z_{i,j}(y) = \frac{f_{i,j} - \langle f_{i,j}^{(CNM)} \rangle}{\sigma_{i,j}^{(CNM)}} , \tag{5.1}$$

where $f_{i,j}$ is the frequency of a pair of journals i and j, $\langle f_{i,j}^{(CNM)} \rangle$ is the expected frequency of observing the same pair but in the configuration null model, and $\sigma_{i,j}^{(CNM)}$ is the standard deviation of $f_{i,j}^{(CNM)}$ in the configuration null model[1]. Note that the z-score for a pair of journals may vary over the years since it

[1]Both the average pair frequency and the corresponding standard deviation are computed over several realizations of the configuration null model, that is several random states of the network that preserve the structural properties mentioned. Each realization is obtained by randomly swapping edges between nodes. See Uzzi et al., 2013 for more details.

is computed for a given publication year. Once they computed how atypical each cited pair was, as defined by the z-score, they assessed the general tendency of a paper to be either conventional or innovative in their citations. Specifically, given a paper, they used the distribution of z-scores for all pairs of journals cited in it, and they took the corresponding median and 10th percentile. The median measures a central tendency toward conventionality, the 10th percentile measures how unconventional the tail of the distribution is. Using these two measures, Uzzi et al., 2013, were able to characterize the relationship between scientific impact and certain innovative practices, showing, for instance, that high-impact papers tend to combine citations to central, conventional papers in the field with citations to uncommon papers.

Here, we adapt this definition of novelty to the case of judicial decisions by making an analogy between the cited papers in a scientific paper and the cited legislation in the text of a judicial decision. With this analogy, we represent the decisions in a corpus and the pattern of citations to law articles as a bipartite network (see Figs. 2.2C and 2.14). Then, taking the sub-graph corresponding to the decisions ruled in a year, we can compute the z-score of each pair of cited law articles adapting the procedure of Uzzi et al., 2013: we first compute the pair frequency observed in the network and, then, the average frequency and standard deviation corresponding to random realizations of the configuration null model[2] to finally use expression 5.1. Once we computed how atypical each pair of law articles is in each year, we compute the novelty of each decision by means of the 10th percentile and the median of the distribution of z-scores of all law article pairs cited in the text:

$$AN_{10th}(d) = -\text{Percentile}_{10}\left(p(z;d)\right) \ , \tag{5.2}$$

$$AN_{med}(d) = -\text{Percentile}_{50}\left(p(z;d)\right) \ , \tag{5.3}$$

where $p(z;d)$ is the z-score distribution function corresponding to the cited law article pairs in a decision d. We added a minus sign to the definition so that positive values of our metrics correspond to positive values of novelty.

[2] In the work by Uzzi et al., 2013, the authors use a configuration null model where the degree of each node is preserved, as well as the publication years of the references cited by each paper. Since we do not have access to the date when each law article became effective, we use a configuration null model that only preserves the degrees of each node.

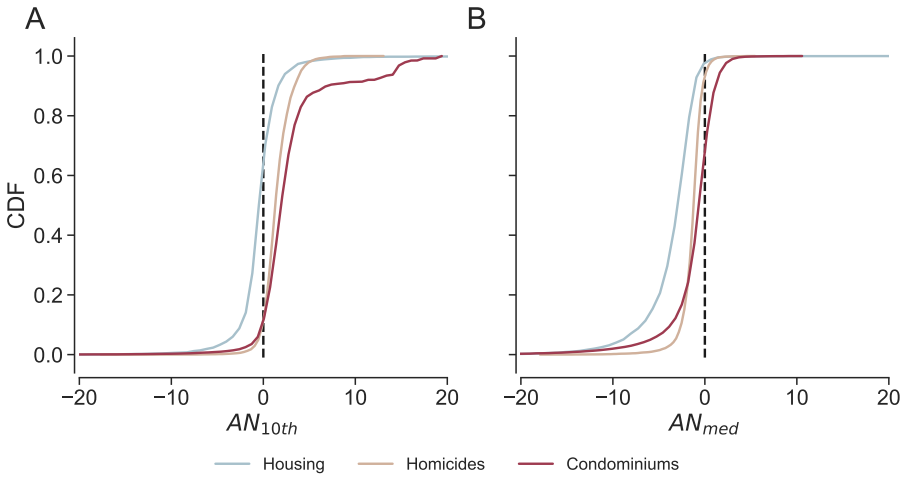


FIGURE 5.1: **Cumulative distribution of atypical novelties function for different corpora.** We show the cumulative distribution function for the 10th percentile (AN_{10th}) (A) and the median (AN_{med}) (B) of the z-score distribution for each judicial decision in each corpora, including Housing, Condominiums and Homicides.

In Fig. 5.1, we show how the values for AN_{10th} and AN_{med} are distributed over all decisions in each corpus. In Fig. 5.1B. we show that in all three cases, decisions have a central tendency to cite legislation more conventional than what is expected by chance. In particular, 97% of decisions in Housing have the median of their z-scores lower than 0, 92% do in Homicides and a 68% do in Condominiums, which in all three cases is considerably lower than the expected 50% that we would get by chance. In the case of the tail of the z-score distribution for each decision, Fig. 5.1A shows that 88% of the decisions both in Homicides and Condominiums have a value for AN_{10th} higher than 0, which is very similar to the expected 90%. In the case of Housing decisions, a 36% of the decisions have a value of AN_{10th} higher than zero, which, in this case, is significantly higher than expected by chance. All in all, results show, first, a clear central tendency for a conventional use of legislation in decisions from all three corpora, and second, a consistent tendency for conventionality higher in decisions from Housing than in the other two corpora.

5.1.2 Novelty as pioneering combinations of cited legislation

The definition of atypical novelty captures the extent to which the citations of legislation by decisions is uncommon compared to that of other contemporary decisions. However, this definition does not capture if a certain use of legislation is uncommon compared to the past, that is, if it is novel in terms of time. For this reason, we propose two definitions of novelty based on detecting pioneering strategies of legislation citation.

To do so, we first have to classify decisions into groups in terms of the use they make of the law. If decisions within a group cite legislation similarly, we can then assign higher values of novelty to earlier decisions in each group. The problem of finding groups of decisions that cite legislation in a similar way, can be formalized as a community detection problem in the bipartite network formed by decisions and the law articles cited (see Fig. 2.2C and 2.14). The proper method to tackle this problem by making the fewest assumptions about the underlying structure of the data is to infer a Stochastic Block Model (SBM) from the mentioned bipartite network (Peixoto, 2019). Note that this problem is equivalent to the one set out in chapters 3 and 4, where we defined legislation topics as groups of law articles in the same bipartite network of decisions and law articles. Whereas there, we were interested in groups of law articles, here, we are interested in groups of decisions. Thus, we take the exact same procedure, that is, we follow an inference approach to find the hierarchical SBM that best fits our data, and we take groups in the lowest level of the hierarchy because it is the most descriptive one (see Fig. 2.14 and section 2.2.3 for more details on the network and the hierarchical stochastic block model inferred).

The stochastic block model corresponds to a partition of the bipartite network of decisions and law articles into non-overlapping groups, with groups constituted only by either decisions or law articles. In what follows, we are only interested in the partition affecting decisions, that is, $\{b_d\}$ is the membership of each decision d to a group b_d, with $d \in [1, D]$, D being the number of decisions. To define novelty using these groups, we consider the ascending chronological order of decisions within a group, and assign to each decision d the corresponding ordinal, I_d. Additionally, we consider the size of group r, n_r, and the correspondence between a decision d and the size of the group where it belongs to: $N(d) = n_r \ : \ r = b_d$. Finally, we define the novelty of a decision d with the following two different definitions, namely GN_1 and GN_2:

$$GN_1(d) = N(d) - I_d \,, \tag{5.4}$$

$$GN_2(d) = \sum_{j=I_d}^{N(d)-I_d} \frac{1}{j} \,. \tag{5.5}$$

Both of these definitions assume that, since we attribute a distinct law citation strategy to each group, those decisions ruled at an earlier time are more novel because they pioneered that citation strategy. In other words, the more recent the decision is within a group, the less novel it is. However, they differ in that, while GN_1 only takes into account the number of decisions succeeding the target decision (one decision in position 28 in a group of size 30 has the same GN_1 as another one in position 1 in a group of size 3), GN_2 balances both the number of preceding and succeeding decisions. In fact, GN_2 can be rewritten in a recursive way in terms of the novelty of the succeeding node as follows:

$$GN_2(d) = \frac{1}{I_d} + GN_2(k) \,, \quad k \,:\, I_k = I_d + 1 \tag{5.6}$$

Both definitions of novelty consider the groups defined by Stochastic Block Model that minimize the description length of the partition given the data (Rissanen, 1978). Nonetheless, finding the solution corresponding to the absolute minimum of the description length (that is, the most probable model given the data) is an NP-hard problem, and therefore, we can only obtain solutions that asymptotically approximate the absolute minimum. This entails that, besides a given model corresponding to a local minimum of the description length that we may use to compute the novelties, there are other models corresponding to other local minima that might indeed have a higher description length but still describe the data similarly. These alternative models usually correspond to partitions with few alterations (that is, they might merge two groups or re-locate a node from one group to another). However, from the point of view of the computation of the novelties, these small reorganizations might translate into significant variations of their value; a node that represents a group by itself in one model but is re-located in a bigger group in an alternative model would change considerably its novelty values. For this reason, we propose to re-define novelties as the average of those defined by expressions 5.4 and 5.5 over several local minima of the minimum

description length:

$$\overline{GN_X}(d) = \frac{1}{R}\sum_k GN_X^{(k)}(d) \tag{5.7}$$

Where $X = \{1,2\}$ and R is the number of different partitions we sample. In this way, the definitions for the group novelties are less sensitive to a specific partition of nodes into groups.

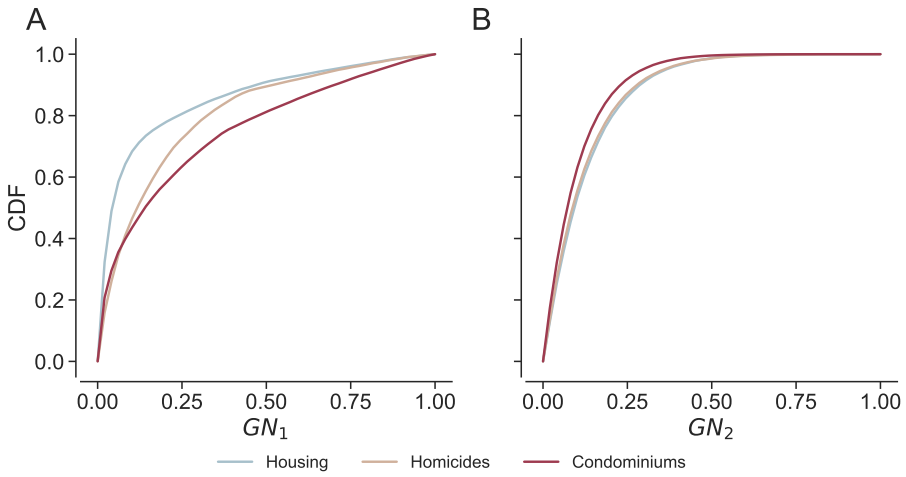


FIGURE 5.2: **Block novelties cumulative distribution function for different corpora.** We show the cumulative distribution function for the Block Novelty 1 (A) and the Block Novelty 2 (B) for each judicial decision in each corpora, namely Housing, Condominiums and Homicides.

In Fig. 5.2, we show the cumulative distribution for both group novelty definitions over the three corpora of decisions. In the case of GN_1, results go in the same direction as the results for the atypical novelty, that is, Housing decisions are distributed towards higher values of novelty. In the case of GN_2, differences between corpora are smaller.

5.1.3 Atypical novelty and group novelty characterize different aspects of innovation

In what follows, we compare the values of the different novelty metrics (AN_{10th}, AN_{med}, GN_1 and GN_2, defined in sections 5.1.1 and 5.1.2) to assess the extent to which they could capture the same aspects of innovation. Figs. 5.3, 5.4

and 5.5 show the correlation between every possible pair of novelty metrics. First, the pair of atypical novelty metrics, AN_{med} and AN_{10th}, and the pair of group novelty metrics, GN_1 and GN_2, show the highest correlation. In the first case, the Spearman's correlation coefficient (C_s) is 0.66 in Housing, 0.65 in Homicides and 0.72 in Condominium, with corresponding p-values being lower than 0.001 in all three cases; in the second, C_s is 0.44 in Housing, 0.57 in Homicides and 0.49 in Condominium, here the corresponding p-values are also lower than 0.001. Second, GN_1 shows a mild positive correlation with both atypical novelty metrics, with C_s scores that go from 0.15 to 0.40 (p-value < 0.001 in all cases). Third, GN_2 shows a small but significantly negative correlation with AN_{10th} only in Condominiums ($C_s = -0.13$, p-value < 0.001) whereas there is no significant correlation in the case of Housing and Homicides ($|C_s| < 0.01$ and p-value > 0.05). The correlation of GN_2 with AN_{med} presents mixed effects, being negative in the case of Homicides and Condominiums ($C_s = -0.07$ and $C_s = -0.10$, respectively) and positive in the case of Housing ($C_s = 0.12$), significant in all three cases with p-values lower than 0.001. All in all, the atypical novelty pair of metrics and the group novelty pair show the smallest values of correlation between metrics outside the pair than within the pair, showing that there are some aspects of innovation that one metric captures that the other does not.

5.1.4 A legal interpretation of innovation

We tested the extent to which the novelty definitions we proposed align with a legal interpretation of innovation. To do so, we asked several legal scholars to evaluate a selection of judicial decisions in terms of the innovation they carry, defined from a legal perspective and taking into account not only the content of the document itself but also the judgment as well. Specifically, an innovative judgment should carry out an interpretation to guarantee the social adaptation of the rule of law. Similarly, the judgment should offer a new solution or interpretation, as long as it is not arbitrary, that solves a social problem or has an impact beyond a particular person.

We performed the test using a selection of decisions from the Housing corpus, including a set of decisions with a particular value of novelty and a set of randomly selected decisions acting as a control group. Specifically, we took the 8 most novel decisions and the 8 least novel decisions for each metric (AN_{10th}, GN_1 and GN_2) and a random selection of 48 decisions from the remaining ones in the corpus. In total, we took $N = 96$ decisions, where half of them

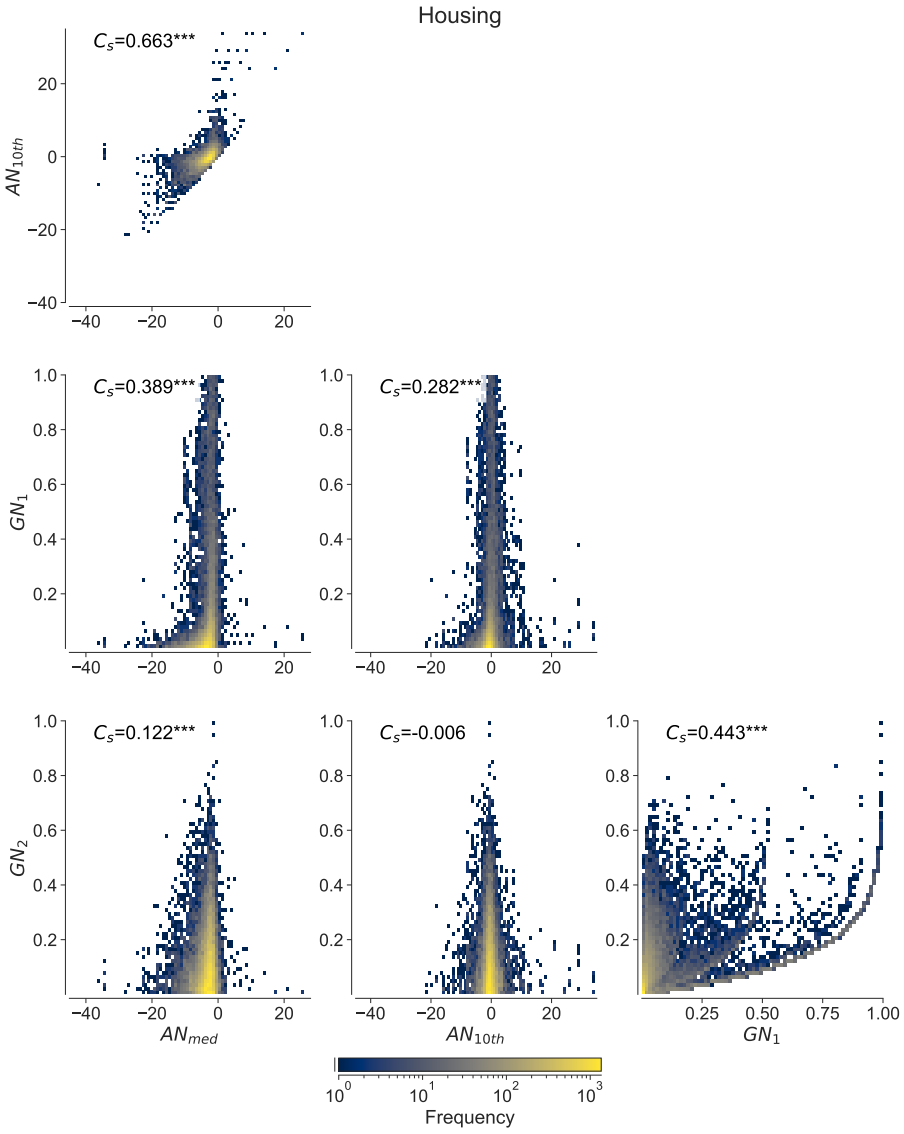


FIGURE 5.3: **Housing corpora novelty metrics pair-wise correlation.** We evaluate the correlation between each pair novelty metrics defined in sections 5.1.1 and 5.1.2: AN_{med}, AN_{10th}, GN_1 and GN_2. For each pair of metrics, we show the Spearman's rank correlation, C_s. Stars indicate when the C_s corresponding p-value is lower than 0.001 (***), lower than 0.01 (**), lower than 0.05 (*) and higher than 0.05 (no stars). We normalized both GN_1 and GN_2 by dividing by the maximum value in the corpus, respectively, so the maximum value is 1 for both metrics.

are particular decisions in terms of novelty and the other half are a random selection. We limited the number of decisions so as to enable a feasible close

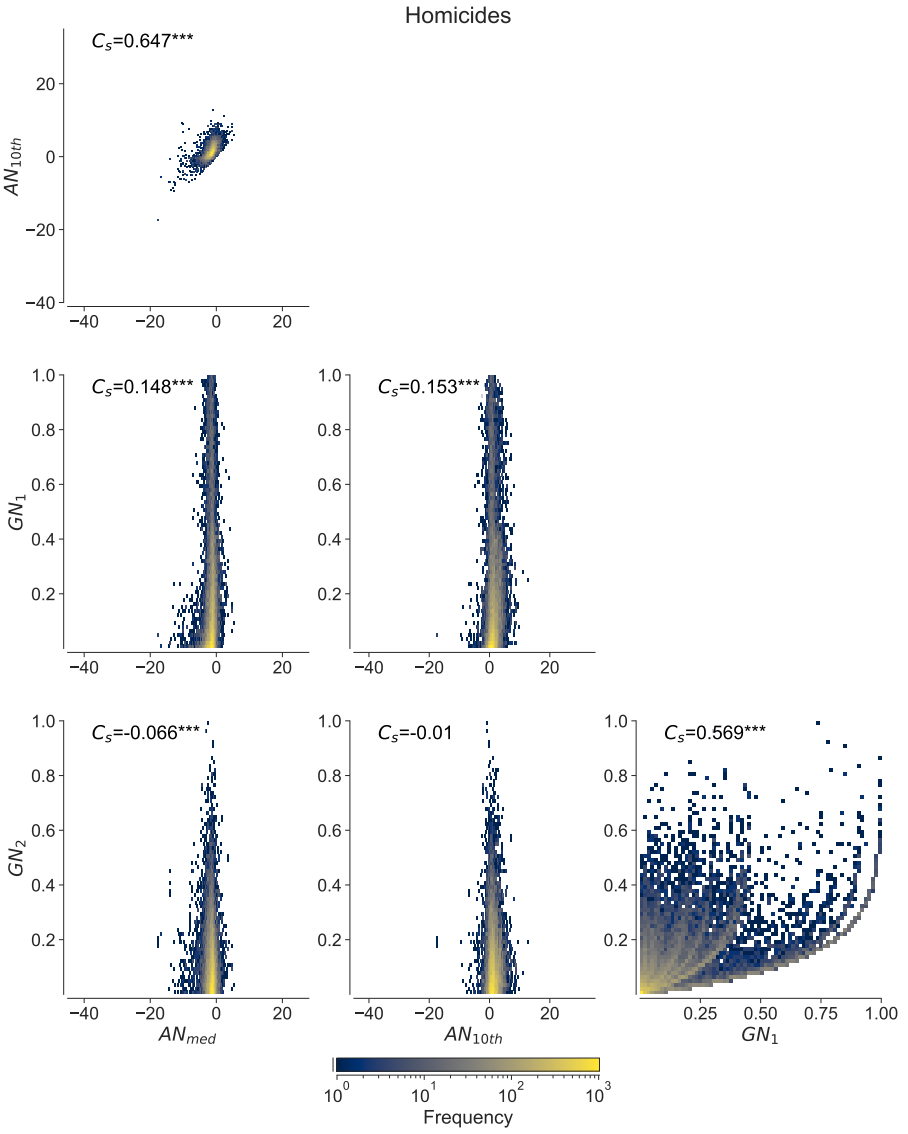


FIGURE 5.4: **Homicides corpora novelty metrics pair-wise correlation.** We evaluate the correlation between each pair novelty metrics defined in sections 5.1.1 and 5.1.2: AN_{med}, AN_{10th}, GN_1 and GN_2. For each pair of metrics, we show the Spearman's rank correlation, C_s. Stars indicate when the C_s corresponding p-value is lower than 0.001 (***), lower than 0.01 (**), lower than 0.05 (*) and higher than 0.05 (no stars). We normalized both GN_1 and GN_2 by dividing by the maximum value in the corpus, respectively, so the maximum value is 1 for both metrics.

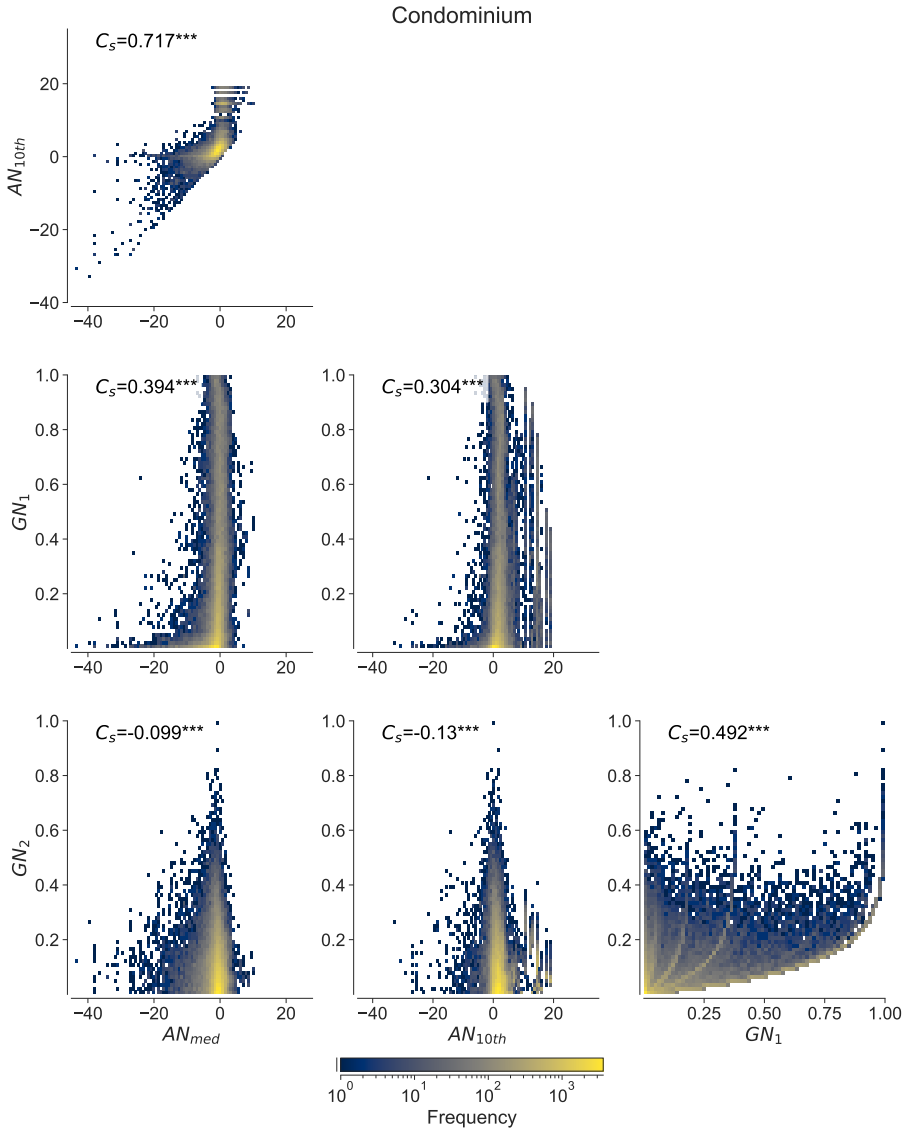


FIGURE 5.5: **Condominium novelty metrics pair-wise correlation.** We evaluate the correlation between each pair novelty metrics defined in sections 5.1.1 and 5.1.2: AN_{med}, AN_{10th}, GN_1 and GN_2. For each pair of metrics, we show the Spearman's rank correlation, C_s. Stars indicate when the C_s corresponding p-value is lower than 0.001 (***), lower than 0.01 (**), lower than 0.05 (*) and higher than 0.05 (no stars). We normalized both GN_1 and GN_2 by dividing by the maximum value in the corpus, respectively, so the maximum value is 1 for both metrics.

reading of all of them. The test was performed by three legal experts, who were asked to make their answers discrete over three possible answers: the

decision being innovative or not, and a neutral answer in between. A list of the top-novelty decisions can be found in Appendix D.

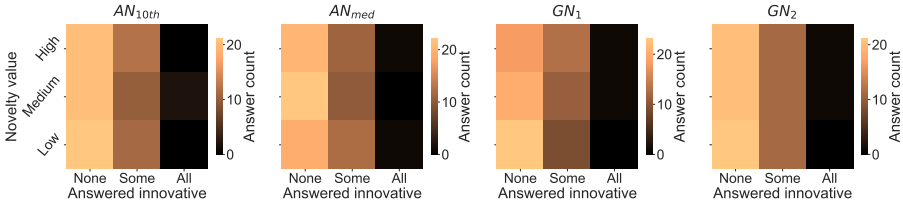


FIGURE 5.6: **Legal interpretation of innovation.** We tested the alignment between the novelty value of our metrics and the innovation value assessed by legal experts. The selection of decisions we tested includes decisions from the Housing corpus, encompassing first, decisions with top novelty values and second, randomly-selected decisions (see section 5.1.4 for more details). We discretized both the novelty values for each definition and the answer given by the legal experts in three boxes each. In the case of the novelty values, we divided values into 3 terciles. In the case of the answers, we divided them into decisions receiving 0, 1 or 2, and 3 votes from experts.

In Fig. 5.6 we show the correlation between the answers and the values of our definitions of novelty. The results do not show enough evidence to conclude that there exists some relationship between any of our novelty metrics and the answers given by legal experts. These results show that the legal aspects considered innovative are not related to the way judges combine legislative sources in the judicial decisions they make.

5.2 Decision impact based on content reuse

Our aim is to define the influence or impact of a given decision on posterior decisions. In science, where the trace for the spread of knowledge can be followed through citations between papers, the impact of a given paper can be estimated from the number of papers that cite it. In the case of judicial decisions, citations from one judicial decision to another exist as well, and thus one could in practice compute the impact of a decision in terms of citations received from posterior decisions. However, measuring the impact of a decision in such terms is not ideal given the nature of our data set for the following reason: In civil law systems, such as the Spanish judicial system, cases are decided on the basis of the law in codes and statutes rather than on the basis of past decisions on similar cases, contrarily to what happens

in common-law systems in the United States or the United Kingdom, for instance. Therefore, in the decisions in our dataset, citations between cases are scarce and most times point out only to a few decisions ruled by the Supreme Court, which represents a very small fraction of the decisions in our data set (see section 2.1.1). Nonetheless, we still assume that judges rely on past decisions when writing and deciding upon current cases and that the trace for such *inspiration* can be found in the reuse of content from one decision to another. For this reason, we alternatively propose to measure the impact of a decision on another one based on the overlap of textual content between the two.

To measure the overlap between decisions, we consider chains of 10 consecutive words in the text, or 10-grams, which allow us to capture the literal reuse of parts of the text. Thus, we represent each decision by the list of 10-grams extracted from the text[3]. Then, having quantified each decision, we measure the overlap between a pair of decisions r and s as the Jaccard index between the corresponding sets of 10-grams, W_d and W_r:

$$J_{dr} = \frac{|W_d \cap W_r|}{|W_d \cup W_r|} \,. \tag{5.8}$$

Then, to measure the impact of a decision, we take into account both the overlap with future decisions and the overlap with past decisions. Specifically, given a decision d ruled in a year $t_d = y$, we compute the average overlap with all $\{D_d^{T^+}\}$ decisions ruled in the posterior T-year time window:

$$D_d^{T^+} = \{\forall s \in \{D\} : t_s \in [t_d + 1, t_d + T]\} \cup \{R_s \neq R_d\} \,, \quad T = 2 \,, \tag{5.9}$$

being $\{D\}$ the set of all decisions in a corpus, and t_s and R_s the year of ruling and the reporting judge of decision s, respectively. Note that we exclude comparisons between same-judge decisions. Similarly, we can consider the corresponding decisions in the past T-year window:

$$D_d^{T^-} = \{\forall s \in \{D\} : t_s \in [t_d - T, t_d - 1]\} \cup \{R_s \neq R_d\} \,. \tag{5.10}$$

[3] To reduce the computational cost of subsequent calculations, we disregard 10-grams only appearing in one decision.

We compute the future and past overlap of a decision d as the average Jaccard index between the decision 10-grams and each of the decisions in the sets defined by expressions 5.9 and 5.10:

$$J_d^{T^+} = \frac{1}{|D_d^{T^+}|} \sum_{s \in D_d^{T^+}} J_{rs} \ , \tag{5.11}$$

$$J_d^{T^-} = \frac{1}{|D_d^{T^-}|} \sum_{s \in D_d^{T^-}} J_{rs} \ . \tag{5.12}$$

Finally, we can compute the impact of a decision as the difference between the overlap with future decisions and the overlap with past decisions:

$$I_r = J_r^{T^+} - J_r^{T^-} \ . \tag{5.13}$$

With this definition of impact, we keep track of the implicit influence that comes with the action of partially reusing the content of other judicial decisions. In this definition, we subtract the future overlap from the past overlap so as to avoid taking into account the impact of those decisions that receive future attention, in part because they are already reusing the content from decisions in the past. Moreover, we restrict the overlap time window to 2 years so that we measure impact with the same conditions for all decisions (in this way, earlier decisions will not have more future decisions than the others, and the last decisions will not have more past decisions than the others). At the same time, we disregard those decisions in the first/last 2-year windows in each corpus.

5.3 Relation between novelty and impact

In what follows, we explore the relationship that might exist between certain novel ways of using legislation and the impact they have on future decisions. To that end, we will, first, compute the correlation between each of the definitions of novelty presented in sections 5.1.1 and 5.1.2 and the impact as defined in section 5.2 and, second, assess the extent to which this correlation is predictive of the impact.

In Fig. 5.7, we show the correlation between each of the novelty metrics and the impact, computed as the Spearman's rank correlation, over the three corpora. Results show that, overall, Housing presents the highest correlation among the three corpora, with a correlation score of 0.32 for GN_1, 0.19 for GN_2, 0.18 for AN_{med} and 0.08 for AN_{10th}, with a corresponding p-value lower than 0.001 in all four cases. In the case of Condominiums, the correlation scores are lower than for Housing, with values below 0.1 for all novelty metrics, having GN_1 and GN_2 the highest correlation scores (the corresponding p-value is lower than 0.001 in all cases except for AN_{med}). In the case of Homicides, the correlation scores fall below 0.1 for all novelty metrics as well, and, moreover, they show opposite values, being -0.08 for AN_{med} and -0.02 for AN_{10th}, whereas 0.04 for GN_1 and 0.1 for GN_1 (the corresponding p-value is lower than 0.001 in all cases except for AN_{med}).

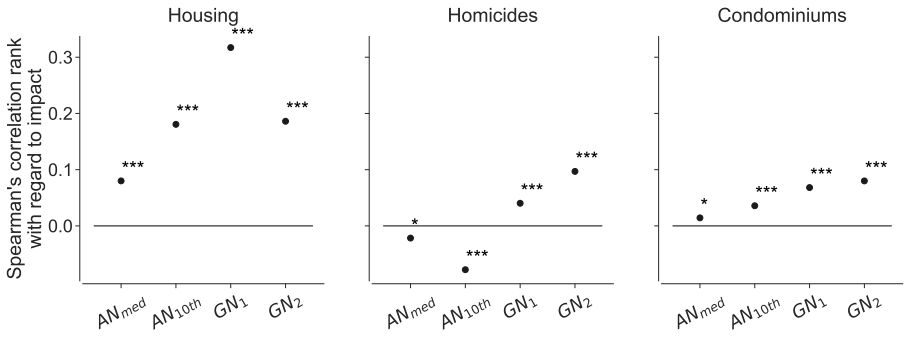


FIGURE 5.7: **Spearman's correlation rank between novelty metrics and impact.** For each corpus, we computed the correlation between the values of impact (defined in section 5.2) and the values of novelties, for each definition of novelty (sections 5.1.1 and 5.1.2) separately. Stars indicate when the p-value is lower than 0.001 (***), lower than 0.01 (**), lower than 0.05 (*) and higher than 0.05 (no stars)

Given these mixed results, we explore if a more complex combination of the novelty metrics can explain the impact of a decision. To do so, we consider the regression problem of finding the dependence between the values of impact (dependent variable) and the combination of novelty values (independent variables), namely, the AN_{med}, AN_{10th}, the GN_1 and the GN_2. Specifically, we train a random forest regression algorithm that, essentially, learns the division of the data (the decisions) in terms of the dependent variables that best explain the independent variable (Hastie et al., 2001). Additionally, to test the regression, we perform a 10-fold cross-validation, that is, we split the data into 10 same-sized parts, we train the regression on 9/10 parts of the

data and test it on the remaining 1/10. For more details on the random forest regression algorithm, see Appendix C. To evaluate the performance, we compute two metrics: the mean squared error (MSE) and the coefficient of determination (R^2). For more details and definitions of metrics, see expressions C.9 and C.10 in Appendix C. Moreover, we compare them with a null model where the values of the dependent variable (the impact) have been randomly shuffled over decisions as a way to calibrate the performance that would be expected by chance.

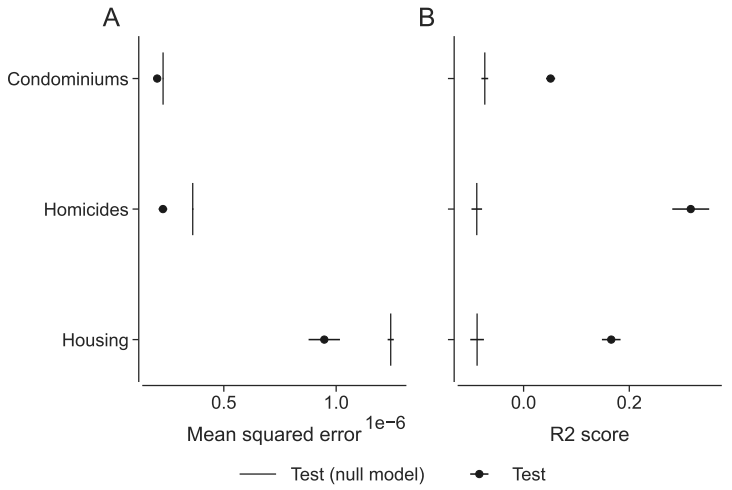


FIGURE 5.8: **Predictive power of decision impact given novelty metrics.** Taking each corpus of decisions, we train a random forest regressor so as to predict the value of the impact of a decision (5.13) given the combination of its novelty values (AN_{10th} and AN_{med}, GN_1 and GN_2). Once trained on part of the data, we test the regressor on the remaining part of the data, repeating the process in a 10-fold cross-validation. We evaluate the performance using two metrics: (A) the mean squared error and, (B), the R squared score (see Appendix C for more details). In both cases, we compare the results to those obtained by randomizing the values of impact over decisions as a null model (horizontal bars in the figure). For the results considering randomized data, error bars represent the standard error over 100 random shuffles of the data. When not visible, error bars are smaller than symbols

The results we show in Fig. 5.8 summarize the forest regression performance over the three corpora of decisions. In short, these results show that the combination of values of novelty of a decision is predictive, to some extent, of the value of the impact of the decision. Specifically, the MSE is lower than expected by chance in all three cases being 88% lower in Condominium, 60% lower in Homicides and 76% in Housing. In the case of the R^2 score, we

obtained a 0.05 score in Condominium, a 0.32 in Homicides and a 0.17 in Housing, while the expected by chance falls below 0 [4].

5.4 Conclusion

In this chapter, we have evaluated the role that innovative behaviors, in terms of the use of legislation, have in judicial decisions. Because innovation is an inherent aspect of science, we have incorporated knowledge from the field of 'Science of Science', where such an aspect has been extensively studied. In our case, we cannot say that innovation is an inherent aspect of judicial sentencing, partially because those behaviors that can be considered out of the ordinary might have a negative connotation (for instance, deciding upon a case without obeying the law). However, these behaviors could still have an important role underlying role in subtle changes that progressively modify how the law is applied over the years. For this reason, we have assessed whether those decisions whose content presents a higher degree of novelty, also gain more attention. We tracked the reuse and content and found that novelty is predictive of later content reuse. We surmised that the reuse of content can be understood as a proxy for the impact decisions have as well as the spread of ideas and knowledge in general over precedents.

Detecting cases in the precedent that present some relevance and escape from those that are trivial or repeated is crucial in the field of legal doctrinal research. In this sense, our work proposes alternative and computational ways of assessing the relevance of a case, which could be useful for not only legal scholars but also legal practitioners. This is important because this detection might be easy to perform by a legal expert through some close examination of cases, but such an evaluation is difficult to conduct as the collection of documents gets larger. Thus, being able to quantify innovative and uncommon practices only from the law articles a judicial decision cites and being able to avoid a close reading of the case entails a significant reduction in time and effort.

[4] While the definition of the R^2 constrains it to range from 0 to 1, in the case where the regression is trained on a specific set of data and then tested on a different, non-overlapping set, the score can be negative. For more details, see Appendix C.

6

Conclusions and perspectives

In this thesis, we laid bare the potential behind the use of large-scale data and computational tools to study documents from courts in the context of legal studies. In particular, one of the general goals of this work has been to analyze certain aspects of the functioning of the judicial system, from a social science perspective, that is, aiming to understand the intricate ways in which agents in the system interact with each other and with external factors as well. At the same time, we used computational and quantitative approaches so as to enable ulterior scrutiny of its functioning based on the evidence.

One of the major points in the development of this thesis has been the access and usability of large databases. To our knowledge, we are the first to engage in a research project that relies on data of such size in the Spanish judicial system, which encompasses approximately 100,000 documents from judicial decisions, corresponding to all appellate courts in the Spanish judicial system, thousands of justices, and spanning more than 20 years. Given these data, a crucial point has been selecting and putting into practice the adequate tools to extract relevant information in a way that scales with the volume of data. To do so, we have relied on natural language processing techniques and network

science representation tools, which have allowed us, for instance, to quantify the content of textual documents while reducing the dimensionality of their representation.

The first aspect of the functioning of the judicial system we studied is related to the time evolution of the content of judicial decisions in a certain domain and, in particular, how these changes over time interplay with legislative changes, the appearance of landmark jurisprudence, and other external societal factors. To study this relationship, we took advantage of the quantification of judicial decisions in terms of both the use of words and the use of legislation, which allowed us, through the use of information-theoretic metrics, to assess the time evolution of the content of the documents. When applied to a housing corpus of decisions, we detected a disrupting period (2015-2016) where a deep reorganization of topics occurred. Importantly, because our methodology not only provides a general quantification of change but also provides an interpretation in terms of the disrupting word and legislation topics responsible for the global changes, we were able to link these changes with important decisions in the precedent from the Court of Justice of the European Union and the Spanish Supreme Court and law modifications related to the criminal offense of squatting. As a result, we have been able to link these disruptions in the content of the documents studied with the consequences of the global financial crisis in 2007. Overall, we showed how using large-scale databases from court documents is an extraordinary source to detect and analyze the most disruptive issues in society.

In the second aspect we studied, we aimed at revealing how the differences in the attributes of reporting judges translate into differences in the way they write their decisions. To do so, we used a spectrum of features that capture from the most structural and stylistic aspects of the writing (namely function-word topics) to those most linked to the reasoning and application of the law (namely content-word topics and legislation topics). By using this spectrum of features, together with the application of machine-learning algorithms for classification purposes, we characterized the differences across different attributes of the reporting judges. First, we revealed that judges are individually very distinguishable. This distinction is mainly caused by the reuse of content judges make from past decisions, which we revealed that it is considerably higher in the case of same-judge decisions than in the case of different-judge decisions. Importantly, these practices concern both the use of words and the use of legislation, showing that the reuse of content is not only a matter of writing style, but also affects the application of the law. Given these

practices, we show that from the word and legislation in a decision, one can predict the exact identity of the reporting judge with high accuracy. Second, we also revealed that there exist differences between women and men that reflect in the content of the decisions they write. These differences are more important when we focus on the non-content and stylistic aspects of the decisions (that is, when using function-word topics), but when we restrict to those aspects related to the law, they are still predictive of the gender (that is, when using content-word topics and legislation topics). Finally, we also revealed that the way judges write decisions is predictive of whether they are early-career judges or they are senior. In this case, the differences rely on both structural and law-related content with similar weight.

The third aspect of the functioning of the judicial system we studied concerns the emergence of innovative behaviors in judicial decisions and how they affect their impact in the future. More specifically, we quantified the degree of innovation that a decision carries from the combinations of cited legislation it makes, and the impact of the diffusion of content it gets in the future.

6.1 Perspectives and future work

The results of this thesis, as well as the methodologies and data we used, are suited to be the starting point for follow-up questions in the field of legal and sociological studies, become potential applications for legal practitioners, or constitute the basis for the promotion of policy interventions.

From a general perspective, our work contributes to the legitimization of quantitative and computational methodologies as well as the use of large-scale data to enable them in a field that still tends to rely on qualitative appraisals to conduct research. In practice, using computational techniques to analyze judicial cases could help conduct reproducibility tests for doctrinal analysis research, contributing to a higher reputation and trustworthiness of the field. Similarly, having conducted successful research using large-scale data encourages the creation (or improvement) of digital repositories that democratize this kind of data. This entails a modification in the protocols of data publication by the corresponding state organizations that goes beyond the simple digitization of documents and incorporates structured, standardized, and machine-readable metadata.

Another contribution of our work comes from the fact that having a better and deeper understanding of the functioning of the judicial system is crucial

to scrutinizing it. In particular, this work shares some similarities with others where the purpose is to detail how the sociological aspects of human interaction, behaviors, and relationships intervene in the function of a social system. We can find a good example of this in the field of science of science or sociology of science, where the examination of the mechanisms underlying the scientific production has enabled a better assessment of scientific impact and success, or has revealed biases and discriminatory attitudes in the scientific publication process related to gender, minorities, etc., to name some examples. In a similar way, in this thesis we have addressed several aspects of the functioning of the judicial system, and the results we obtained open the door for legal scholars and other stack-holders to address and remedy those aspects that require improvement.

A remarkable contribution from our analysis of the time evolution of the content of judicial decisions has been the use of interpretable, in a legally-oriented way, computational tools. Specifically, we went beyond the use of classical topic models to quantify the use of words in a document by proposing a legislation topic model that models the cited articles in the law. Thus, our results in terms of disruptive periods and topics have been far more meaningful than they would have been by using common topic models, revealing non-trivial relations among law articles and understanding better the role of specific legislation in judicial sentencing, for instance, and thus becoming a proof of concept of its usefulness. Additionally, our methodology allows for expanding the analysis to other judicial systems from other countries, other legal fields, other courts, other time periods, etc., that could present very different characteristics of the data (language, size, etc.) without requiring any modification. Specifically, we devised a general approach to estimate the KL divergence in a precise and efficient way that prevents the biases caused by sparse samples of the data. As a result, our methodology can be applied independently of the size of the data or the characteristics of the topic model used to quantify the documents.

Our results from the analysis of how the differences in judges attributes translate into differences in the way they write their judicial decisions show that these differences go beyond simple stylistic and individual fingerprints and also lay on the legal-related content substantial to the application of the law. This result opens the door to evaluating in more detail if these differences could be behind biases and a malfunction of the judicial system that obstructs the administration of justice. To do so, it would be crucial to have access to systematic data on the verdict of the case, where there is no consensus on

whether there is a clear effect from the attributes of the judge, as well as access to data from other fields and time periods. Moreover, although our methodology is able to produce results that allow us to discern whether the differences lie in the style or structure of documents, or in aspects related to the law, they are still not reducible to a small and easily explainable set of elements, which suggests that these differences are inherently complex. However, more effort on feature extraction from documents to reveal differences could foster a clear interpretation of them. All in all, our results open the door to analyzing in more detail the effects on the administration of justice that might come from a malfunctioning case allocation among judges, opening the door to revising these protocols.

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

A

Bayesian estimation of the Kullback-Leibler Divergence for sparsely sampled distributions

In this appendix, we expose how to estimate the Kullback-Leibler Divergence (D_{KL}) when the distributions are obtained from sparse data using a hierarchical Bayesian approach that provides an efficient, precise and general estimation of the metric. The content of this appendix is based on the paper by Piga et al., 2023.

The D_{KL}, as other information theoretical metrics such as the Shannon entropy, is measured on distributions. These distributions, in practice, have to be estimated from experimental observations. However, this inference process is difficult for many real complex systems since, due to experimental limitations, the observations are often sparse, and statistical estimates of these distributions the functions on them such as the D_{KL} can be severely biased (Levina et al., 2022).

In this appendix, we focus on the particular case of discrete (or categorical) distributions, $\boldsymbol{\rho}$, with ρ_i, $i = 1, \ldots, K$, where K is the number of possible states (or categories), which is known and fixed. In this particular case, inferences about $\boldsymbol{\rho}$ and any function must be based on n_i, the number of observations in the i-th state (with $N = \sum_i n_i$ the sample size). In the undersampled regime, that is, when $N \lesssim K$, the challenge is thus, from the sparse observations $\{n_i\}$, to infer the probability ρ_i of each category i and finally estimate the D_{KL}. However, since the D_{KL} is a functional applied on two probability distributions, we have two actually estimate both of them first. The D_{KL} is defined as follows:

$$D_{\mathrm{KL}}(\boldsymbol{\rho}\|\boldsymbol{\sigma}) = \sum_{i=1}^{K} \rho_i \log_2 \frac{\rho_i}{\sigma_i} \,, \tag{A.1}$$

A theoretically well-founded approach to tackle this problem is provided by the principles of conditional probability, encapsulated in Bayes' theorem (Jaynes, 2003). This framework is in general preferable because of its transparency —it requires that all assumptions of the underlying generative model for the data are made explicit, expressed via the choice of a likelihood function and a prior distribution that reflects the knowledge about the system before observing any data. In probabilistic reasoning, the combination of observations and prior distribution provides an updated (posterior) probability distribution of the quantity under study. Other estimation strategies make implicit assumptions and often provide only point estimates, as opposed to full distributions.

In a very similar problem, namely trying to estimate the Shannon entropy from sparse samples, the state of the art is the approach by Nemenman et al., 2001, which assumed a mixture of Dirichlet priors as the generative model for the categorical distribution, obtaining a precise estimator of the Shannon entropy that works for a wide variety of distributions, even in the sparse sampling regime $N \lesssim K$ (Nemenman et al., 2001 and Nemenman et al., 2004). However, the estimator does not provide estimates for the distribution $\boldsymbol{\rho}$, which limits its applicability to only information theoretic quantities that can be expressed in terms of entropies, such as the mutual information and the Jensen-Shannon distance, but not for the D_{KL}.

The approach we expose in this appendix is a semi-analytical estimator grounded in probabilistic considerations and without any *ad hoc* assumptions. In particular, we consider Dirichlet generative models and we use a hierarchical

Bayesian approach to extract as much information as possible from the few observations at hand. When applied to the case of the Shannon entropy, we can estimate the expected value and higher order moments with precision at least comparable to the NSB estimator, and most often better (See Piga et al., 2023, for more details). Additionally, because our method provides estimates of the probability distribution, it can be used to obtain accurate estimations of the Kullback-Leibler divergence. In this case, our approach also performs equally or better than existing estimators.

In general, to estimate the function $\mathcal{F}(\boldsymbol{\rho})$ of $\boldsymbol{\rho}$, where $\boldsymbol{\rho}$ is a discrete distribution function, $\boldsymbol{\rho} = \{\rho_i;\, i = 1, \dots, K\}$ with $\sum_i \rho_i = 1$, that can be estimated from some observations on each estate $\mathbf{n} = \{n_i;\, i = 1, \dots, K\}$, with $\sum_i n_i = N$ independent observations, one can compute the posterior distribution of $\mathcal{F}(\boldsymbol{\rho})$ given these observations:

$$p\left(\mathcal{F}|\mathbf{n}\right) = \int d\boldsymbol{\rho}\; \delta\left(\mathcal{F} - \mathcal{F}(\boldsymbol{\rho})\right) p(\boldsymbol{\rho}|\mathbf{n}) \,, \tag{A.2}$$

where $p(\boldsymbol{\rho}|\mathbf{n})$ is the posterior of the distribution $\boldsymbol{\rho}$ given the counts $\mathbf{n}$. Using the laws of conditional probability, we can write this posterior as:

$$p(\boldsymbol{\rho}|\mathbf{n}) = \frac{p(\mathbf{n}|\boldsymbol{\rho})\, p(\boldsymbol{\rho})}{p(\mathbf{n})} \,, \tag{A.3}$$

where $p(\mathbf{n}|\boldsymbol{\rho})$ is the likelihood, $p(\boldsymbol{\rho})$ is the prior over distributions, and $p(\mathbf{n}) = \int d\boldsymbol{\rho}\; p(\mathbf{n}|\boldsymbol{\rho})\; p(\boldsymbol{\rho})$ is the evidence and acts as normalization factor. The likelihood is the probability of the empirical observations $\mathbf{n}$ given $\boldsymbol{\rho}$; for independent multinomial samples, the probability of observing an event of type i is ρ_i, and the full likelihood is the following:

$$p(\mathbf{n}|\boldsymbol{\rho}) = N! \prod_{i=1}^{K} \frac{\rho_i^{n_i}}{n_i!} \,. \tag{A.4}$$

The prior $p(\boldsymbol{\rho})$ expresses the probability of each distribution $\boldsymbol{\rho}$ prior to observing any data. Symmetric Dirichlet distributions are convenient priors because they are a generative model for a broad class of discrete distributions. They are parameterized as follows:

$$p(\boldsymbol{\rho}|\beta) = \frac{1}{\mathrm{B}_K(\beta)} \prod_{i=1}^{K} \rho_i^{\beta-1} \,, \quad \mathrm{B}_K(\beta) = \frac{\Gamma(\beta)^K}{\Gamma(\beta K)} \,, \tag{A.5}$$

where Γ is the gamma function, while the hyperparameter β is a real, positive number known as the concentration parameter.

Then, by setting our prior $p(\boldsymbol{\rho}) = p(\boldsymbol{\rho}|\beta)$, we can compute the posterior distribution for $p(\boldsymbol{\rho}|\mathbf{n}, \beta)$ in A.3 and the posterior for $\mathcal{F}(\boldsymbol{\rho})$ in A.2. The expected value of this posterior $\langle \mathcal{F} \rangle = \int d\mathcal{F}\, \mathcal{F}\, p(\mathcal{F}|\mathbf{n})$ minimizes the mean-squared error (Wolpert and Wolf, 1995), and its mode is a consistent estimator, meaning that it converges to the true value of $\mathcal{F}(\boldsymbol{\rho})$ when the number of observations increases, regardless of the prior and, in particular, regardless of the hyperparameter β. However, for very scarce samples the posterior $p(\mathcal{F}|\mathbf{n})$ is dominated by the prior, which makes the choice of beta very sensitive. For instance, in the case of estimating the entropy of $\boldsymbol{\rho}$, that is, when $\mathcal{F}(\boldsymbol{\rho}) = S(\boldsymbol{\rho})$, Nemenman et al., 2001, noticed that S is narrowly determined by, and monotonically dependent on, β. Nemenman et al., 2001, and Nemenman et al., 2004, overcame this situation by proposing a prior proportional to an infinite mixture of Dirichlet priors, as a way to obtain a prior completely agnostic over the entropies. However, in the case of the estimation of the D_{KL}, the function suffers from two differences compared to the estimation of S that prevents us from applying an equivalent approach. First, D_{KL} is not a combination of the Shannon entropies, oppositely to other information theoretical functions such as mutual information or the Jensen-Shannon distance. Second, the D_{KL} is unbounded, which makes any attempt to find a hyperprior in the spirit of Nemenman et al., 2001, and Nemenman et al., 2004 to result in an improper hyperprior.

In the following, we address these limitations with a hierarchical Bayes point estimate for β that will provide an accurate estimation of the probability distribution and the subsequent estimation of D_{KL}.

A.1 A hierarchical Bayes point estimate for β

Here, we posit that the success of the NSB approach stems, not from mixing infinitely many values of the concentration parameter β, but rather from the flexibility to accommodate for *any particular value* of β. Indeed, we surmise that, in general, only a narrow interval of β values are compatible with a given observation $\mathbf{n}$ and therefore contribute to the mixture, whereas most others do not contribute. Motivated by this, we propose an approach that aims to directly estimate the value of β that most contributes to the posterior given the data $\mathbf{n}$.

First, we observe that the posterior $p(\boldsymbol{\rho}|\mathbf{n})$ can be written as

$$\begin{aligned} p(\boldsymbol{\rho}|\mathbf{n}) &= \int d\beta\, p(\boldsymbol{\rho}|\mathbf{n},\beta)\, p(\beta|\mathbf{n}) \\ &= \int d\beta \frac{p(\mathbf{n}|\boldsymbol{\rho})\, p(\boldsymbol{\rho}|\beta)}{p(\mathbf{n}|\beta)}\, p(\beta|\mathbf{n})\ , \end{aligned} \tag{A.6}$$

where we have applied Bayes' rule, and the fact that $\mathbf{n}$ conditioned on $\boldsymbol{\rho}$ is independent of β, so that $p(\mathbf{n}|\boldsymbol{\rho},\beta) = p(\mathbf{n}|\rho)$. Then, we assume that the conditional distribution $p(\beta|\mathbf{n})$ is very peaked around a given value $\beta^\star$, so that the posterior $p(\boldsymbol{\rho}|\mathbf{n})$ can be approximated as

$$p(\boldsymbol{\rho}|\mathbf{n}) \approx \frac{p(\mathbf{n}|\boldsymbol{\rho})\, p(\boldsymbol{\rho}|\beta^\star)}{p(\mathbf{n}|\beta^\star)}\ . \tag{A.7}$$

This approximation, sometimes referred to as *empirical Bayes*, is a point estimate for the fully hierarchical probabilistic model given by $p(\mathbf{n}|\boldsymbol{\rho})$ and $p(\boldsymbol{\rho}|\beta)$. Eq. (A.7) is identical to Eq. (A.3), with the difference that the concentration parameter is now the most likely value of β given the observed counts $\mathbf{n}$, that is,

$$\beta^\star = \underset{\beta}{\operatorname{argmax}}\ p(\beta|\mathbf{n}) = \underset{\beta}{\operatorname{argmax}}\ \frac{p(\mathbf{n}|\beta)\, p(\beta)}{p(\mathbf{n})}\ , \tag{A.8}$$

where $p(\mathbf{n}|\beta) = \int d\boldsymbol{\rho}\, p(\mathbf{n}|\beta,\boldsymbol{\rho}) p(\boldsymbol{\rho}|\beta)$.

Our goal here is to obtain the expression for $p(\beta|\mathbf{n})$ and maximize it to obtain $\beta^\star$. Applying the laws of conditional probability, we have:

$$p(\beta|\mathbf{n}) = \frac{p(\beta)}{p(\mathbf{n})} p(\mathbf{n}|\beta)\ , \quad p(\mathbf{n}|\beta) = \int d\boldsymbol{\rho}\, p(\mathbf{n}|\beta,\boldsymbol{\rho}) p(\boldsymbol{\rho}|\beta)\ . \tag{A.9}$$

The terms in expression A.9 are the following: $\mathrm{p}(\mathbf{n})$ is the evidence and acts as a normalization constant, thus not relevant for a maximization operation. $p(\mathbf{n}|\beta,\boldsymbol{\rho})$ does not depend on β and it is given by equation A.4. $p(\boldsymbol{\rho}|\beta)$ is the Dirichlet prior, given by A.5. Finally, the hyperprior $p(\beta)$ reflects our prior knowledge about the shape of the distribution of β. To be completely agnostic in this regard, we can use a uniform hyperprior:

$$p_U(\beta) = \frac{1}{\Delta\beta} = \text{const.}\ , \quad \Delta\beta = \beta_{\max} - \beta_{\min}\ , \tag{A.10}$$

with cut-offs $0 < \beta_{\min} < \beta_{\max} < \infty$. Thus, we can express $p(\beta|\mathbf{n})$ as follows:

$$p(\beta|\mathbf{n}) = \frac{p(\beta)}{p(\mathbf{n})}\frac{\Gamma(\beta K)}{[\Gamma(\beta)]^K}\int d\boldsymbol{\rho}\,\delta\left(\sum_{i=0}^{K}\rho_i - 1\right) N!\prod_{i=1}^{K}\frac{\rho_i^{n_i+\beta-1}}{n_i!} \quad (A.11)$$

$$= \frac{p(\beta)}{p(\mathbf{n})}\frac{\Gamma(\beta K)}{[\Gamma(\beta)]^K}\frac{N!}{\prod_{i=1}^{K} n_i!} I \quad (A.12)$$

The integral I in the equation A.12 has to be computed in the $K-1$ simplex, where the probability distribution $\boldsymbol{\rho}$ is normalized, that is, $\sum_{i=1}^{K}\rho_i = 1$. Then, one factor in I (for the sake of convenience we take the last, $i = K$) can be written as $\rho_K = \sum_{i=1}^{K-1}\rho_i$, and the product in the integral takes the form:

$$\prod_{i=1}^{K}\rho_i^{n_i+\beta-1} = \rho_1^{n_1+\beta-1}\cdots\rho_{K-1}^{n_{K-1}+\beta-1}\left[1-(\rho_1+\cdots+\rho_{K-1})\right]^{n_K+\beta-1}. \quad (A.13)$$

Therefore, the integral over the ρ_{k-1} variable can be expressed as follows:

$$I_{K-1} = \int_0^{1-\sum_{i=0}^{K-2}\rho_i} d\rho_{K-1}\,\rho_{K-1}^{n_{K-1}+\beta-1}\left(1-\rho_{K-1}-\sum_{i=1}^{K-1}\rho_i\right)^{n_K+\beta-1}, \quad (A.14)$$

where the limits of integration respect the normalization constraint. The integral I_{K-1} can be evaluated by knowing that:

$$\int_0^{1-R} d\rho\rho^a\left[1-(\rho+R)\right]^b = \frac{\Gamma(a+1)\Gamma(b+1)(1-R)^{a+b+1}}{\Gamma(a+b+2)} \quad (A.15)$$

Thus,

$$I_{K-1} = \frac{\Gamma(n_{K-1}+\beta)\Gamma(n_K+\beta)}{\Gamma(n_K+n_{K-1}+2\beta)}\left[1-\sum_{i=1}^{K-2}\rho_i\right]^{n_K+n_{K-1}+2\beta-1} \quad (A.16)$$

Similarly, for the integral over the ρ_{K-2} variable, we have:

$$I_{K-2} = \frac{\Gamma(n_{K-2}+\beta)\Gamma(n_K+n_{K-1}+2\beta)}{\Gamma(n_K+n_{K-1}+n_{K-2}+3\beta)}\left[1-\sum_{i=1}^{K-3}\rho_i\right]^{n_K+n_{K-1}+n_{K-2}+3\beta-1} \quad (A.17)$$

We note that the term $n_K + n_{K-1} + 2\beta$ cancels out since being both in the denominator of A.16 and in the numerator of A.17. Then, considering this act for all the integrals, we can finally express I as:

$$I = \frac{\prod_{i=1}^{K} \Gamma(n_i + \beta)}{\Gamma(N + K\beta)} \,. \tag{A.18}$$

Then, we can add I to the expression A.12:

$$p(\beta|\mathbf{n}) = \frac{p(\beta)}{p(\mathbf{n})} \frac{\Gamma(\beta K)}{[\Gamma(\beta)]^K} \frac{N!}{\prod_{i=1}^{K} n_i!} \frac{\prod_{i=1}^{K} \Gamma(n_i + \beta)}{\Gamma(N + K\beta)} \,. \tag{A.19}$$

Once we have the expression for $p(\beta|\mathbf{n})$, we can now maximize it to find the corresponding value of β, $\beta^\star$. First, we take the logarithm:

$$\begin{aligned} \log p(\beta|\mathbf{n}) = \log p(\beta) - \log p(\mathbf{n}) + \log N! - \sum_{i=1}^{K} \log n_i! + \log \Gamma(K\beta) \\ -K \log \Gamma(\beta) + \sum_{i=1}^{K} \log \Gamma(n_i + \beta) - \log \Gamma(N + K\beta) \end{aligned} \tag{A.20}$$

We want to find $\beta^\star$ that satisfies:

$$\left. \frac{d \log p(\beta|\mathbf{n})}{d\beta} \right|_{\beta=\beta^\star} = 0 \,. \tag{A.21}$$

Then, deriving the expression A.20:

$$\begin{aligned} \frac{d \log p(\beta|\mathbf{n})}{d\beta} = -K \frac{d}{d\beta} \log \Gamma(\beta) + \frac{d}{d\beta} \log \Gamma(K\beta) + \sum_{i=1}^{K} \frac{d}{d\beta} \log \Gamma(n_i + \beta) \\ - \frac{d}{d\beta} \log \Gamma(N + K\beta) \,. \end{aligned} \tag{A.22}$$

Where we have considered that $p(\beta)$ is uniform and therefore the derivative is zero. The equation A.22 can be expressed as follows:

$$\frac{d \log p(\beta|\mathbf{n})}{d\beta} = -K\psi_0(\beta) + K\psi_0(K\beta) + \sum_{i=1}^{K} \psi_0(n_i + \beta) - K\psi_0(N + K\beta) \,. \tag{A.23}$$

where $\psi_0(z)$ is the digamma function and we have used that $d\log\Gamma(z)/dz = \psi_0(z)$. Knowing that $\psi_0(z+n) = \sum_{m=0}^{n-1}\frac{1}{z+m} + \psi(z)$:

$$\frac{d\log p(\beta|\mathbf{n})}{d\beta} = \sum_{i=1}^{K}\sum_{m=0}^{n_i-1}\frac{1}{m+\beta} - \sum_{m=0}^{N-1}\frac{K}{m+K\beta} \tag{A.24}$$

Therefore, to maximize $p(\beta|\mathbf{n})$, $\beta^\star$ has to fulfill the following condition:

$$\sum_{i=1}^{K}\sum_{m=0}^{n_i-1}\frac{1}{m+\beta^\star} - \sum_{m=0}^{N-1}\frac{K}{m+K\beta^\star} = 0\,. \tag{A.25}$$

Considering both distributions that generated by a Dirichlet prior and distributions *atypical*, (that is, they cannot be attributed to or have a negligible probability of being generated from a symmetric Dirichlet prior) such as multi-modal distributions, Zipf distributions or distributions with added structural zeros[1].

[1]Generated by first drawing a distribution from a symmetric Dirichlet prior with a given β and second, forcing half of the categories to have zero probability.

B

Word and legislation topics

In this appendix, we show the content of a selection of topics in terms of either their words (in the case of word topics) or law articles (in the case of legislation topics). The selection of topics encompass those analyzed in Chapter 3 for being the most disruptive ones: word topics 108 and 14 (at hierarchical level, $hl = 2$) and legislation topics 41 and 42 ($hl = 1$).

We performed the legal interpretation of these topics (see section 3.3) based on the words and legislation that appear in them (see tables B.1, B.2, B.3, B.4 and B.5). Moreover, the tables show the hierarchical structure that topics have, illustrating how topics in a given hierarchical level merge with others at the hierarchical level above and, similarly, break down in smaller sub-topics in the hierarchical levels below. Thus, topics at higher hierarchical levels tend to be more general while they are more specific at lower levels, offering different levels of detail for the corresponding legal interpretation.

Topic tables provide the weight of each word/law article within the topic, as provided by the topic model (see section 2.2.3) and that represents the fraction

of times the word/law article appears in relation to others within the topic; that is, all weights sum up to 1 within the topic.

hl=3, topic 16

Weight (%)	Word
	hl=2, topic 124
3.83	consumidor
1.45	consumidores
0.47	incorporación
0.39	general_defensa_consumidores
0.37	usuarios
0.36	consumidores_usuarios
0.31	lcgc
0.24	nulas
0.18	empresarios
0.18	lgdcu
0.16	declarada_nula
0.13	artículo_lcgc
0.13	consumidora
0.12	acción_colectiva
0.11	contratos_consumidores
0.1	cláusula_declarada_nula
0.09	condición_consumidores
0.08	disposición_adicional_primera
	hl=2, topic 108
[...]	[...]
	hl=2, topic 123
3.31	préstamo
2.05	prestatario
1.81	hipoteca
1.17	préstamo_hipotecario
0.93	préstamo_garantía_hipotecaria
0.88	artículo_ley_hipotecaria
0.8	prestamista
0.75	préstamos
0.64	prestatarios
0.62	entidad_financiera
0.6	hipotecaria
0.57	préstamos_hipotecarios
0.55	garantía_hipotecaria
0.49	contrato_préstamo
0.48	parte_prestataria
0.41	escritura_préstamo_hipotecario
0.39	amortización
0.34	hipotecario
0.32	crédito_hipotecario
0.29	prestataria
0.29	hipotecas
0.29	entidad_prestamista
0.28	comisiones
0.28	ley_hipotecaria
0.22	entidad_crédito
0.17	adquisición_vivienda
0.16	escritura_préstamo
0.15	capital_prestado
0.14	capital_pendiente
0.12	ley_marzo
0.11	notarios
0.09	acreedor_hipotecario
0.09	trlgcu
0.09	cuotas_amortización
0.08	pública_préstamo_hipotecario
0.08	banco_prestamista
0.08	cláusula_litigiosa
0.08	parte_deudora
	hl=2, topic 72
1.76	deudor
1.1	acreedor
1.09	ejecutante
0.93	ejecución_hipotecaria
0.68	procedimiento_ejecución_hipotecaria
0.65	despacho_ejecución
0.61	subasta
0.45	parte_ejecutante
0.39	ejecutados
0.34	parte_ejecutada
0.34	título_ejecutivo
0.33	proceso_ejecución
0.32	deudas
0.3	oposición_ejecución
0.29	proceso_ejecución_hipotecaria
0.27	deudores
0.26	procedimiento_ejecución
0.23	sobreseimiento_ejecución
0.22	remate
0.22	garantía_real
0.2	entidad_ejecutante
0.19	demanda_ejecutiva
0.19	hipotecante
0.18	ley_mayo
0.18	finca_hipotecada
0.17	deudor_hipotecario
0.17	valor_tasación
0.15	protección_deudores_hipotecarios
0.15	bien_hipotecado
0.15	deudora
0.13	bienes_hipotecados
0.12	incidente_oposición
0.11	entidad_acreedora
0.1	procedimiento_hipotecario
0.1	intereses_costas
0.09	existencia_cláusulas_abusivas
0.08	totalidad_deuda
0.08	título_judicial

TABLE B.1: **Words in topic 16, hierarchical level (hl) 3.** We show all words constituting topic 16 at $hl = 3$, with the corresponding weight (see main text in the present Appendix B). We display words organized in lower-level topics (hl=2, topics 124, 108, 123 and 72). We omitted the words in topic 108, (hl=2, in green) since we display them in table B.2, showing the hierarchical organization of these words at subsequent lower levels.

hl=2, topic 108

Weight (%)	Word
	hl=1, topic 376
6.33	cláusula
3.6	cláusulas
	hl=1, topic 393
2.46	directiva
2.14	abusiva
0.5	artículo_directiva
0.48	abusivas_contratos_celebrados
0.32	cee_consejo_abril
0.29	derecho_interno
	hl=1, topic 386
2.43	vencimiento_anticipado
1.83	cláusula_vencimiento_anticipado
0.56	cláusulas_vencimiento_anticipado
0.43	anticipadamente
0.42	duración_cuantía_préstamo
0.39	capital_intereses
0.38	vencimiento_anticipado_préstamo
0.34	medios_adecuados_eficaces
0.26	carácter_suficientemente_grave
0.26	poner_remedio_efectos
0.25	resolución_anticipada
0.23	cuotas_mensuales
0.22	fundamento_ejecución
0.21	concurra_justa_causa
0.21	cuotas_impagadas
0.21	verdadera_manifiesta_dejación
0.21	obligaciones_carácter_esencial
0.18	cláusula_sexta_bis
0.18	sola_cuota
0.17	sexta_bis
0.16	incumplimiento_grave
0.15	dar_vencido
0.14	relativa_vencimiento_anticipado
	hl=1, topic 377
[...]	[...]
	hl=1, topic 421
1.89	cláusula_suelo
1.68	tipo_interés
1.46	transparencia
1.04	objeto_principal_contrato
0.87	cláusulas_suelo
0.86	control_transparencia
0.67	interés_variable
0.52	euribor
0.41	falta_transparencia
0.41	oferta_vinculante
0.37	diferencial
0.36	nominal_anual
0.36	índice_referencia
0.35	tipo_referencia
0.33	tipo_interés_variable
0.33	devolución_cantidades
0.33	clara_comprensible
0.32	variación_tipo_interés
0.31	intereses_ordinarios
0.29	nulidad_cláusula_suelo
0.26	condiciones_financieras
0.24	variabilidad
0.24	incorporación_contrato
0.22	aplicación_dicha_cláusula
0.21	sencillez
0.2	carga_económica
0.2	interés_fijo
0.2	permita_consumidor
0.19	oscilaciones
0.19	om
0.18	cláusula_tercera_bis
0.18	comprensibilidad_real
0.17	cláusulas_techo
0.17	doble_control_transparencia
0.16	tipo_mínimo
0.16	tercera_bis
0.16	denominada_cláusula_suelo
0.16	aplicación_cláusula_suelo
0.16	elementos_esenciales_contrato
0.15	préstamo_interés_variable
0.15	control_inclusión
0.14	cantidades_cobradas
0.14	sacrificio_patrimonial
0.14	recalcular
0.13	podrá_ser_inferior
	hl=1, topic 303
1.68	tjue
0.93	abusivo
0.67	tribunal_justicia
0.64	eu_c
0.55	clausula
0.41	jurisprudencia_tjue
0.37	derecho_unión
0.34	declaración_abusividad
0.33	litigio_principal
0.29	banco_español_crédito
0.25	todas_consecuencias_oportunas
0.25	unicaja_banco
0.22	vía_ejecutiva
0.22	largo_plazo
0.21	clausulas
0.21	véase_sentencia
0.21	llegado_aplicarse_opone
0.2	cuestiones_prejudiciales
0.19	tribunal_nacional
0.19	debe_interpretarse_juez
0.19	permite_resolución
0.19	derecho_comunitario
0.18	consumidor_profesional_circunstancia
0.18	equilibrio_real
0.18	cuestión_prejudicial_planteada
0.17	cláusula_contrato_celebrado
0.17	constatado_carácter_abusivo
0.17	artículos_apartado
0.17	obiter_dicta
0.16	disposición_supletoria_derecho
0.15	anular_contrato
0.14	artículo_apartado_propia
	hl=1, topic 409
1.46	intereses_demora
1.29	interés_demora
1.16	intereses_moratorios
0.65	interés_remuneratorio
0.61	interés_legal_dinero
0.57	interés_moratorio
0.53	devengo
0.41	intereses_remuneratorios
0.35	moderación
0.31	dos_puntos
0.29	abusivos
0.26	recálculo
0.23	tres_veces
0.23	interés_remuneratorio_pactado
0.23	cláusula_intereses_moratorios
0.21	adquisición_vivienda_habitual
0.2	remuneratorio
0.15	desproporcionadamente_alta
0.12	calculados
	hl=1, topic 395
1.19	empresario
1.11	condiciones_generales
1.02	condiciones_generales_contratación
0.81	desequilibrio
0.75	adherente
0.68	condición_general
0.56	texto_refundido_ley
0.54	condición_general_contratación
0.52	consumidor_usuario
0.45	control_abusividad
0.41	real_decreto_legislativo
0.41	exigencias_buena_fe
0.4	desequilibrio_importante
0.35	negociación_individual
0.35	directiva_cee
0.33	predisponente
0.27	negociada_individualmente
0.26	perjuicio_consumidor
0.26	consecuencias_económicas
0.25	retribución
0.23	control_contenido
0.22	bienes_servicios
0.21	negociado_individualmente
0.21	cláusulas_negociadas
0.19	control_incorporación
0.18	concertados_consumidores
0.18	trlgdcu
0.16	artículo_directiva_cee
0.16	artículo_trlgdcu
0.15	cláusulas_predispuestas

hl=1, topic 377

Weight (%)	Word
	hl=0, topic 821
8.04	abusividad
2.36	abusividad_cláusula
2.0	sentencia_tjue_marzo
	hl=0, topic 585
7.08	carácter_abusivo_cláusula
1.93	derecho_nacional
0.8	aziz
	hl=0, topic 794
6.98	cláusulas_abusivas
0.78	sentencia_tjue_junio
0.6	moderar
	hl=0, topic 684
5.94	cláusula_abusiva
2.37	tal_cláusula
1.9	aplicación_cláusula
	hl=0, topic 683
5.19	juez_nacional
3.06	artículo_apartado_directiva
1.3	efecto_disuasorio
0.9	modificar_contenido
0.75	normas_nacionales
0.72	juez_nacional_debe
0.66	contrato_celebrado_profesional
0.62	juez_nacional_facultad
	hl=0, topic 860
4.73	dicha_cláusula
1.06	cláusula_cuestión
0.87	referida_cláusula
0.83	declarada_abusiva
	hl=0, topic 556
4.57	abusivas
1.6	dichas_cláusulas
1.22	predispuesta
0.65	cláusula_impugnada
	hl=0, topic 615
4.37	carácter_abusivo
2.85	contrato_préstamo_hipotecario
	hl=0, topic 689
3.81	nulidad_cláusula
2.58	declaración_nulidad_cláusula
	hl=0, topic 586
3.4	derechos_obligaciones_partes
1.05	momento_celebración
0.94	detrimento_consumidor_desequilibrio
0.87	naturaleza_bienes_servicios
0.67	apreciará
	hl=0, topic 624
3.35	sentencia_tjue
2.57	cláusula_contractual
1.65	carácter_abusivo_cláusulas
1.65	cláusulas_contractuales
0.73	contratos_préstamo

TABLE B.2: **Words in topic 108, hierarchical level (hl) 2, and topic 377, hl=1.** We show all words constituting topic 108 at $hl = 2$ (in green) and topic 377 at $hl = 1$ (in orange) with the corresponding weight (see main text in the present Appendix B). Topic 108 merges with others at higher-level topic 16 (hl=3, see table B.1). We display words in topic 108 organized in lower-level topics (hl=1, topics 376, 393, 386, 377, 421, 303, 409 and 395). We omitted words in topic 377, (hl=1, orange) and we display them in the right column, where we show the hierarchical organization of words at the lowest level, hl=0.

hl=3, topic 1

Weight (%)	Word
	hl=2, topic 89
3.37	pena
2.9	delitos
0.11	pena_imponer
	hl=2, topic 128
3.19	ministerio_fiscal
0.65	delictiva
0.61	cometido
0.46	comisión_delito
0.3	cometer
0.29	cometidos
0.24	protegido
0.22	hechos_delictivos
0.21	sala_segunda
0.18	delictivo
0.16	delictivas
0.15	delincuente
0.15	ilícitos
0.14	persecución
0.14	presuntamente
0.12	impunidad
0.1	utilizó
0.1	cometer_delito
0.09	delito_grave
	hl=2, topic 14
[...]	[...]
	hl=2, topic 95
2.03	policía
1.94	juicio_oral
1.72	plenario
1.5	agentes
0.94	policial
0.56	atestado
0.42	policías
0.41	encontró
0.33	vista_oral
0.31	comisaría
0.3	fase_instrucción
0.3	agentes_policía
0.25	policía_nacional
0.21	sumarial
0.2	agentes_policiales
0.18	funcionarios_policiales
0.16	cuerpo_nacional_policía
0.16	guardia
0.14	actuantes
0.14	sumariales
0.12	sesiones
0.12	atestado_policial
0.1	multa_euros
0.09	cometiendo
0.08	inmediatez
0.08	preconstituida
0.07	funcionarios_policía
	hl=2, topic 24
1.82	código_penal
1.64	autor
1.15	acusación_particular
0.74	atenuante
0.57	acusaciones
0.57	intimidad
0.53	espacio
0.49	inviolabilidad_domicilio
0.39	confesión
0.36	eximente_incompleta
0.35	perjudicados
0.33	individuo
0.33	atenuante_analógica
0.3	violenta
0.3	leve
0.26	atenuación
0.23	lugares
0.21	privacidad
0.21	criminalmente_responsable_delito
0.21	cualificada
0.21	circunstancia_atenuante
0.19	incluidas_acusación_particular
0.16	atenuantes
0.15	vida_privada
0.13	recinto
0.13	reparación_daño
0.12	imagen
0.11	atenuante_cualificada
0.11	analógica
0.11	eximente_completa
	hl=2, topic 93
1.31	diligencias
0.96	instrucción
0.6	falsedad
0.52	estafa
0.37	delito_continuado
0.28	continuado
0.19	engaño
0.18	falso
0.16	falsa
0.15	apropiación
	hl=2, topic 23
1.26	prueba_cargo
1.2	convicción
1.02	presunción_inocencia
0.95	credibilidad
0.91	versión
0.68	infracción_ley
0.67	sala_segunda_tribunal
0.64	responsabilidad_civil
0.5	tribunal_sentenciador
0.49	violencia_intimidación
0.46	relato
0.4	contradicciones
0.39	hecho_probado
0.35	situado_calle_número
0.35	ministerio
0.31	razonabilidad
0.22	acto_plenario
0.22	quince_días
0.22	subsunción
0.2	punitivo
0.19	descargo
0.18	autor_criminalmente_responsable
0.18	canon
0.15	segunda_sentencia
0.13	artículo_código_penal
0.13	narración
0.11	fallo_condenatorio
0.11	antijurídica
0.1	inferencias
0.1	acervo_probatorio
0.1	control_casacional
0.1	relato_histórico
0.1	corroboraciones
0.1	denuncia_vulneración
	hl=2, topic 87
1.21	investigación
0.91	teléfono
0.81	detención
0.78	detenido
0.49	policiales
0.4	funcionarios
0.38	móvil
0.37	funcionario
0.32	conclusiones_definitivas
0.3	teléfono_móvil
0.29	blanca
0.28	sospechas
0.24	responsable_concepto_autor
0.23	policía_judicial
0.22	vigilancias
0.22	constitutivos_delito
0.21	oficio_policial
0.2	teléfonos
0.2	conclusiones_provisionales
0.2	detenidos
0.2	jefe
0.17	mostró
0.17	sospecha
0.17	previsto_penado_artículo
0.17	investigaciones
0.16	piezas
0.15	descubrimiento
0.14	concepto_autor
0.13	habilitante
0.11	entradas
0.11	impusiera
0.11	carnet
0.11	hechos_objeto_acusación
0.1	compañeros
0.1	detener
0.1	detenida
0.09	hechos_relatados
0.09	implicación
0.09	momento_detención
	hl=2, topic 20
0.95	imputado
0.58	diligencias_previas
0.5	juez_instrucción
0.5	instructor
0.39	imputados
0.37	hechos_denunciados
0.35	investigado
0.29	querella
0.24	sobreseimiento_provisional
0.22	recurso_reforma
0.21	investigados
0.21	inculpado
0.17	juez_instructor
0.16	instructora
0.15	diligencias_investigación
0.14	procesamiento
0.12	sobreseimiento_libre
0.11	criminalidad
0.11	sobreseimiento_provisional_archivo
0.11	sala_penal
0.1	investigar
	hl=2, topic 1
0.92	guardia_civil
0.86	autores
0.52	bien_jurídico
0.47	eximente
0.39	bienes_jurídicos
0.35	situación_necesidad
0.32	amenaza
0.3	penado
0.26	norma_penal
0.26	policía_local
0.25	punible
0.23	móviles
0.22	identificados
0.2	mossos_esquadra
0.2	autoridad_judicial
0.2	injusto
0.14	condenas
0.13	ejecutoria
0.13	acción_típica
0.13	ofendido
0.11	trabajos_beneficio_comunidad
0.11	programa
0.1	desvalor
	hl=2, topic 113
0.53	procedimiento_abreviado
0.52	relato_fáctico
0.47	cuota_diaria_euros
0.43	indebida_aplicación
0.34	principio_acusatorio
0.31	meses_cuota_diaria
0.28	aplicación_indebida
0.26	dni
0.25	defenderse
0.24	objeto_acusación
0.22	escrito_acusación
0.22	apertura_juicio_oral
0.21	seis_euros
0.19	pena_meses
0.19	acusa
0.18	transformación
0.15	motivo_puede_prosperar
0.15	dni_número
0.12	euros_cuota_diaria
0.1	acusación_pública
0.1	euros_día
0.09	número_antecedentes_penales
0.09	multa_meses
0.09	escrito_defensa
0.08	acusación_formulada

TABLE B.3: **Words in topic 1, hierarchical level (hl) 3.** We show all words constituting topic 1 at $hl = 3$, with the corresponding weight (see main text in the present Appendix B). We display words organized in lower-level topics (hl=2, topics 89, 128, 14, 95, 24, 23, 87, 20, 1, and 113). We omitted the words in topic 14, (hl=2, in green) since we display them in table B.4, showing the hierarchical organization of these words at subsequent lower levels.

hl=2, topic 14	
Weight (%)	**Word**
——— hl=1, topic 20 ———	
18.98	denunciante
12.77	denunciado
——— hl=1, topic 14 ———	
11.93	denunciada
11.91	denunciados
2.06	juicio_faltas
2.0	entidad_denunciante
1.49	delitos_leves
1.31	denunciadas
——— hl=1, topic 468 ———	
[...]	[...]

hl=1, topic 468	
Weight (%)	**Word**
——— hl=0, topic 778 ———	
28.24	bien_jurídico_protegido
——— hl=0, topic 1052 ———	
20.28	derecho_penal
13.43	principio_intervención_mínima
3.99	intervención_derecho_penal
——— hl=0, topic 1169 ———	
7.39	juzgado_penal
——— hl=0, topic 1067 ———	
5.83	precepto_penal
5.42	perturbación_posesión
5.1	protección_penal
3.62	riesgo_posesión
3.59	intervención_penal
3.11	objeto_material

TABLE B.4: **Words in topic 14, hierarchical level (hl) 2, and topic 468, hl=1.** We show all words constituting topic 14 at $hl = 2$ (in green) and topic 468 at $hl = 1$ (in orange) with the corresponding weight (see main text in the present Appendix B). Topic 14 merges with others at higher-level topic 1 (hl=3, see table B.3). We display words in topic 14 organized in lower-level topics (hl=1, topics 20, 14 and 468). We omitted words in topic 468, (hl=1, orange) and we display them in the right column, where we show the hierarchical organization of words at the lowest level, hl=0.

hl=2, topic 24

Weight (%)	Law article
	hl=1, topic 41
[...]	[...]
	hl=1, topic 42
[...]	[...]
	hl=1, topic 40
9.35	Artículo 790 de la Ley de Enjuiciamiento Criminal
4.21	Artículo 792 de la Ley de Enjuiciamiento Criminal
2.86	Artículo 973 de la Ley de Enjuiciamiento Criminal
1.69	Artículo 791 de la Ley de Enjuiciamiento Criminal
1.08	Artículo 131 del Código Penal
0.86	Artículo 967 de la Ley de Enjuiciamiento Criminal
0.69	Artículo 118 de la Ley de Enjuiciamiento Criminal
0.68	Artículo 13 del Código Penal
0.57	Artículo 229 de la Ley Orgánica del Poder Judicial.
0.53	Artículo 132 del Código Penal
0.53	Artículo 255 del Código Penal
0.5	Artículo 246 del Código Penal
0.32	Artículo 969 de la Ley de Enjuiciamiento Criminal
0.28	Artículo 130 del Código Penal
0.28	Artículo 37 de la Ley Orgánica de protección de la seguridad ciudadana
0.22	Artículo 803 de la Ley de Enjuiciamiento Criminal
0.21	Artículo 37 del Código Penal
0.21	Artículo 981 de la Ley de Enjuiciamiento Criminal
0.21	Artículo 964 de la Ley de Enjuiciamiento Criminal
0.18	Artículo 971 de la Ley de Enjuiciamiento Criminal
0.17	Artículo 963 de la Ley de Enjuiciamiento Criminal
0.16	Artículo 319 del Código Penal
0.12	Artículo 247 del Código Penal
0.12	Artículo 739 de la Ley de Enjuiciamiento Criminal
0.11	Artículo 245 de la Ley Orgánica del Poder Judicial.
0.11	Artículo 635 del Código Penal
0.1	Artículo 531 del Código Penal
0.09	Artículo 965 de la Ley de Enjuiciamiento Criminal
0.09	Artículo 105 de la Ley de Enjuiciamiento Criminal
0.09	Artículo 624 del Código Penal
0.08	Artículo 975 de la Ley de Enjuiciamiento Criminal
0.08	Artículo 972 de la Ley de Enjuiciamiento Criminal
0.08	Artículo 14 del Convenio de Roma
0.07	Artículo 133 del Código Penal
0.07	Artículo 134 del Código Penal
0.07	Artículo 970 de la Ley de Enjuiciamiento Criminal
0.07	Artículo 966 de la Ley de Enjuiciamiento Criminal
0.06	Artículo 33 Otras decisiones del Juez de Menores de la Ley Penal del Menor
0.06	Artículo 334 del Código civil
0.06	Artículo 145 del Código Penal
0.06	Artículo 968 de la Ley de Enjuiciamiento Criminal
0.06	Artículo 11 del Convenio de Roma
0.05	Artículo 334 del Código Penal
0.05	Artículo 32 de la Ley Orgánica de protección de la seguridad ciudadana
0.05	Artículo único. Modificación de la ley de enjuiciamiento criminal. de la Ley Orgánica 13/2015, de 5 de octubre, de modificación de la Ley de Enjuiciamiento Criminal para el fortalecimiento de las garantías procesales y la regulación de las medidas de investigación tecnológica.
0.05	Artículo 441 del Código Penal
0.05	Artículo 974 de la Ley de Enjuiciamiento Criminal
0.05	Artículo 13 de la Ley Orgánica del Tribunal Constitucional
0.04	Artículo 668 de la Ley de Enjuiciamiento Criminal
0.04	Artículo 225 del Código Penal
0.04	Artículo 557 del Código Penal
0.04	Artículo 41 de la Ley del Patrimonio de las Administraciones Públicas.
0.03	Artículo 8 Principio Acusatorio de la Ley Penal del Menor
0.03	Real Decreto 1955/2000, de 1 de diciembre, por el que se regulan las actividades de transporte, distribución, comercialización, suministros y procedimientos de autorización de instalaciones de energía eléctrica: Artículo 87. Otras causas de la suspensión del suministro.
0.03	Artículo 215 de la Ley de Enjuiciamiento Criminal
0.03	Artículo 438 del Código Penal
0.02	Artículo 383 del Código Penal
0.02	Artículo 311 del Código Penal
0.02	Artículo 4. Derechos de las víctimas y de los perjudicados de la Ley Penal del Menor
0.02	Disposición derogatoria única. Derogación normativa. de la Ley Orgánica de protección de la seguridad ciudadana
0.02	Disposición transitoria cuarta. Juicios de faltas en tramitación. de la Ley Orgánica 1/2015, de 30 de marzo, por la que se modifica la Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal.
0.02	Artículo 608 del Código Civil
0.02	Artículo 389 del Código civil
0.02	Artículo 37 de la Ley 4/2015, de 17 de junio, de mejora de la estructura territorial agraria de Galicia
0.01	Artículo 503 del Código Penal
0.01	Artículo 792 del Código civil
0.01	Artículo 38 de la Ley Orgánica del Tribunal Constitucional
0.01	Artículo 558 del Código Penal de 1973
0.01	Artículo 562 del Código Penal de 1973
0.01	Artículo 45 de la Ley Penal del Menor
0.01	Artículo 40 de la Convención sobre los Derechos del Niño
0.01	Artículo 33 de la Ley de Enjuiciamiento Criminal
0.01	Ley 9/1999, de 13 de mayo, Ordenación del Territorio de Canarias: Artículo 62. Derechos y deberes de los propietarios de suelo rústico.
0.01	Artículo 927 de la Ley de Enjuiciamiento Criminal
0.01	Artículo 615 del Código Civil
0.01	Artículo 564 del Código Civil
0.01	Ley 2/1974 de 13 de febrero (Jefatura), sobre Colegios Profesionales: Artículo 9.
0.01	Artículo 213 del Código Penal
0.01	Artículo 541 del Código Civil

hl=1, topic 41

Weight (%)	Law article
	hl=1, topic 63
96.14	Artículo 245 del Código Penal
2.32	Artículo 962 de la Ley de Enjuiciamiento Criminal
1.13	Artículo 430 del Código civil
0.24	Artículo 425 del Código Penal
0.18	Artículo 956 de la Ley de Enjuiciamiento Civil de 1881.

hl=1, topic 42

Weight (%)	Law article
	hl=1, topic 64
38.81	Artículo 82 de la Ley Orgánica del Poder Judicial.
14.97	Artículo 976 de la Ley de Enjuiciamiento Criminal
8.35	Artículo 977 de la Ley de Enjuiciamiento Criminal
2.28	Artículo 205 del Código Penal
	hl=1, topic 160
15.08	Artículo 795 de la Ley de Enjuiciamiento Criminal
4.38	Artículo 239 del Código Penal
3.44	Artículo 796 de la Ley de Enjuiciamiento Criminal
0.69	Artículo 543 de la Ley Orgánica del Poder Judicial.
0.69	Artículo 82 del Código Penal
0.69	Artículo 5 de la ley orgánica del Tribunal Constitucional.
0.4	Artículo 203 de la Ley Orgánica del Poder Judicial.
0.18	Artículo 18 de la Ley del Patrimonio de las Administraciones Públicas.
0.14	Artículo 29 de la Ley del Patrimonio de las Administraciones Públicas.
0.07	Artículo 273 del Código Penal
	hl=1, topic 171
8.57	Artículo 248 del Código Penal
1.27	Artículo 252 del Código Penal

TABLE B.5: **Law articles in topic 24, hierarchical level (hl) 2.** We show all law articles constituting topic 24 at $hl = 3$, with the corresponding weight (see main text in the present Appendix B). We display law articles organized in lower-level topics (hl=1, topics 41, 42 and 40). We omitted the words in topic 41, (hl=1, in blue) and topic 42, (hl=1, in red) since we display them in the right column, showing the hierarchical organization of these law articles at subsequent lower levels.

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

C

Random forest algorithm

A random forest algorithm is a supervised learning algorithm used for classification and regression problems (Breiman, 2001). In this appendix, we will summarize some of its main characteristics and explain its uses over the course of the present thesis.

Classification and regression Random forests normally address classification or regression problems. In both of them, some data is used to learn the relation between some independent variables and a dependent variable with the purpose of predicting or extrapolating the value of the dependent variable when new values for the independent variables are given. Specifically, we can define the data set as follows:

$$\mathcal{D}^{(N)} = \{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \dots, (\mathbf{x_N}, y_N)\}, \tag{C.1}$$

where N denotes the size of the data set (the number of data points), and the independent variables are represented as vectors in linear space, that is,

$\{\mathbf{x_i}\} \in \mathbb{R}^p$, p being the number of features in the data set. When the dependent variable y is continuous, we talk about a regression problem, when y is a categorical variable, we talk about a classification problem.

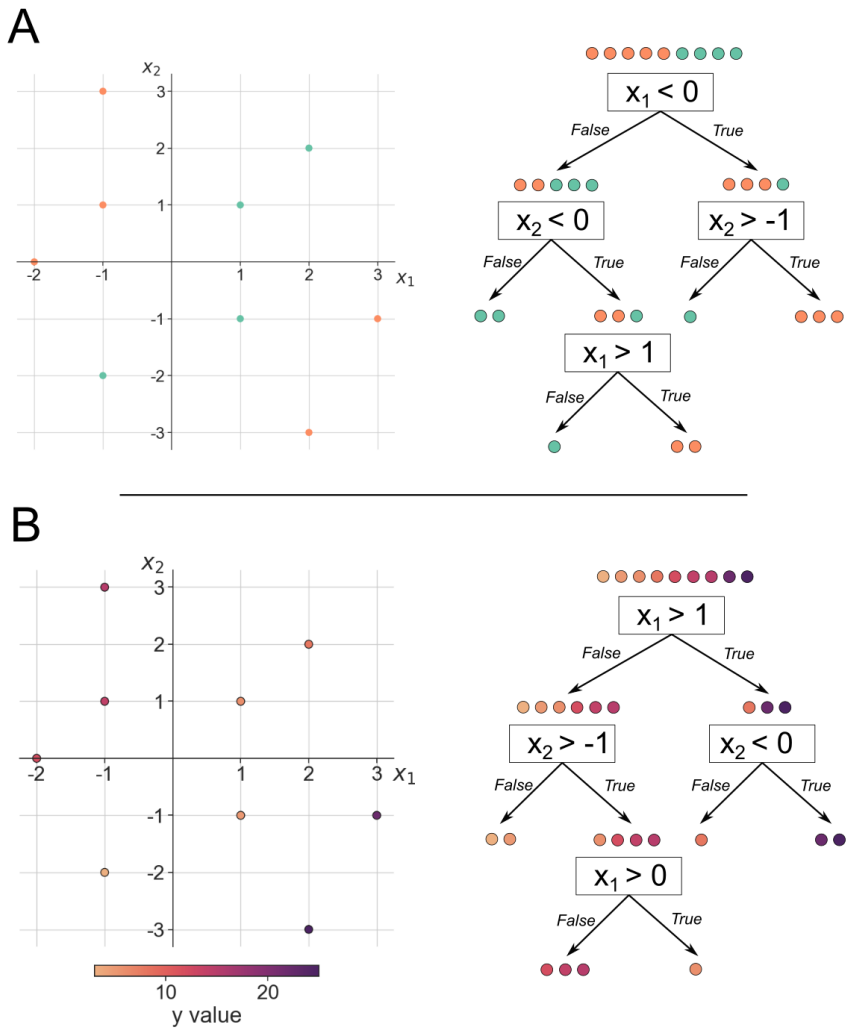


FIGURE C.1: **Decision tree for classification and regression problems.** The data set consist of 9 points; for each point a categorical variable y (orange/green) and two features, x_1 and x_2. The decision tree divides the data points according certain rules over the features x_1 and x_2 until each division contains pure groups of data, that is, either green or orange points.

C.1 Definition of Random Forest

Decision trees A random forest algorithm is built upon many simple decision trees. A decision tree is a binary tree that recursively splits the data points according to some conditions over the values of the features. In classification problems, the tree splits the data until each node has pure group of data points, that is, all categories of the points in the group are the same (see Fig. C.1A). In regression problems, the tree splits the data until a fixed maximum depth of the tree is reached (see Fig. C.1B). In classification problems, decision trees set the splitting conditions so that they maximize, over features and feature values, the information gain after splitting the data. The information gain is equivalent to the reduction in impurity, computed as the entropy of the fraction of categories within a group. Specifically, and for the binary case where y can belong to two categories, the impurity of a group of data points is the following:

$$I(w) = w \log w + (1 - w) \log(1 - w) \,, \tag{C.2}$$

where w is the fraction of points with the values of y belonging to one of the two categories. Then, the reduction of impurity between a group and the two resulting from a split:

$$\Delta I = I_s - I(w_0) \,, \tag{C.3}$$

$$I_s = wI(w_a) + (1 - w)I(w_b) \,, \tag{C.4}$$

where w_0, w_a and w_b are the fraction of points corresponding to the first category in the original group, the first split and the second split, respectively. In the case of a regression problem decision trees seek to reduce the impurity but, this time, computed as the variance of the data. Finally, one estimates the value of y' associated to an unseen given features, $\mathbf{x}'$ by going over the ruled defined by the tree applying them to $\mathbf{x}'$ and taking the predicted value as the estimate of y'.

A decision tree is a greedy algorithm: it tends to classify very well the data (that is, it presents very low bias), but, at the same time, it presents a high variability, that is, a slight change on the values of the features would result in significant changes on the tree. In other words, decision trees have a tendency

to over fit and therefore they are not accurate when extrapolating new data. Random forests aggregate results from many different trees with the purpose of reducing the variability of the results while keeping the low bias provided by each tree. Essentially, random forest achieve this by the use of bootstrap sampling, model aggregation and random feature selection.

Bootstrap sampling This technique, mostly used in random forests but that can be used in any statistical classification and regression algorithm, is used to create multiple alternative data sets from the original data. Taking a data set defined as in expression C.1, a bootstrap sampling consists of creating K alternative data sets with size $m < N$ by randomly sampling, and with replacement, data points from the original data set $\mathcal{D}$.

Model aggregation Once these K alternative data set have been sampled, a decision tree is trained on each of them, hence the concept of forest from the use of an ensemble of trees. Then, to provide the estimate of an unseen data point, $\mathbf{x}$, the random forest algorithm considers the majority vote over the outcomes of all trees, in the case of a classification problem, and the average outcome in the case of a regression problem, hence the concept of aggregation.

The use of bootstrap sampling together with model aggregation has two main effects on the estimates. First, since the data used to fit each tree is identically distributed, the expectation for an estimate is the same for each of them and for the average as well, which results in that the aggregated result has the same bias as a single tree would have. Second, the variance is reduced as the number of trees increases, limited by the effect of correlation between the trees. For more details, see Hastie et al., 2001. Finally, to reduce the correlation between the trees and thus improve the reduction of the variance, random forest use a random selection of features for each tree. Specifically, when training a tree, the algorithm selects first a subset of features randomly and trains the tree using only these selected features. This technique has the main purpose of de-correlating the trees from each other. Several studies have found that the recommended number of selected features is $m = \sqrt{p}$ for classification problems and $m = p/3$ for regression problems XXX.

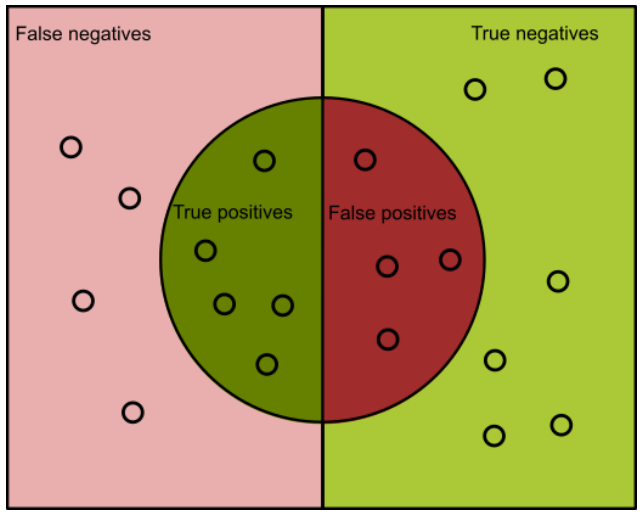


FIGURE C.2: **Binary classification performance scheme.** Each dot represents a classification made in a binary classification, where data can be classified as either 'positive' or 'negative'. Dots inside the circle are those points predicted as positive, while those outside the circle are those points predicted as negative. Dots on the left side (both inside and outside the circle) are those points that are actually positive, while dots on the right side (both inside and outside the circle) are those that are actually negative. The green colors represent those points classified correctly while the red colors represent those classified incorrectly. Then, the data can be separated in four regions: true positives (predicted positive correctly), false positives (predicted positive but they are actually negative), false negatives (predicted negative but they are actually positive) and true negatives (predicted negative correctly).

C.2 Performance metrics

Different metrics can be used to evaluate the performance of the predicted results of any statistical learning algorithm. Here, we summarize those used in this thesis, separating between those used for classification problems and those used for regression problems.

C.2.1 Classification

See Fig. C.2 for scheme on the predicted values in a binary classification problem, including the concepts of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

Accuracy The ratio of data points correctly classified.

$$accuracy = \frac{TP + TN}{P + N} \tag{C.5}$$

Precision The ratio between the number of true positives and the total number of predicted positive data points.

$$precision = \frac{TP}{TP + FP} \tag{C.6}$$

Recall The ratio between the number of true positives and the total number of positive data points.

$$recall = \frac{TP}{TP + FN} \tag{C.7}$$

F_1 score The harmonic mean between the precision and the recall.

$$F_1 = 2\frac{precision \cdot recall}{precision + recall} = \frac{2TP}{2TP + FP + fn} \tag{C.8}$$

Area under the receiver operating curve The receiver operating curve (ROC) illustrates the ability of a classifier to discern between classes in a binary classification. Specifically, it plots the true positive rate (TP/P) over the false positive rate (FP/N) for the whole range of possible values of the threshold, that is, the value used in the algorithm to separate between positive and negative predictions. The area under the ROC is used to evaluate the performance of a classifier. It is defined between 0 and 1, where 0.5 corresponds to the performance of a random guesser and 1, to a perfect classifier.

C.2.2 Regression

In regression problems, performance metrics measure the 'distance' between the predicted values and the real ones.

Mean squared error (MSE) The MSE computes the average squared distance between the predicted values $\{f_i\}$ and the actual values $\{y_i\}$.

$$MSE = \frac{1}{N}\sum_{i}^{N}(y_i - f_i)^2 \tag{C.9}$$

Coefficient of determination Also denoted by R^2, evaluates the performance computing the error in the prediction as a fraction over the total variation of the data. It is expressed as 1 minus the mentioned fraction. A value of 1 implies a perfect regression, a value of and 0 implies a prediction error the size of the variation of the data (on average).

$$R^2 = 1 - \frac{\sum_i^N (y_i - f_i)^2}{\sum_i^N (y_i - \overline{y})^2} \tag{C.10}$$

C.3 Cross-validation testing

Cross-validating a statistical learning algorithm is a way to test a model, that has been trained using some part of the data, on the remaining part of the data. In this way, the models are tested with 'out-of-sample' data. The procedure for a k-fold cross-validation works as follows. First, the data is split into k equally-sized portions. Second a model is trained using $k-1$ portions of the data and tested on the k remaining portion of the data. Recursively, the model is trained and tested using all other $k-1$ combinations of data. Thus, the performance is evaluated k times, one for each portion of the data tested. In classification problems, the splits of the data are chosen so they keep the overall proportion of categories.

Leave-one-out cross validation In the extreme case, one could set k so it matches the size of the data, $k = N$. In this case, each data point is tested with a model that has been trained using all the remaining data points. This cross-validation scheme is suited for small data sets as it is more computationally expensive.

UNIVERSITAT ROVIRA I VIRGILI
QUANTITATIVE LARGE-SCALE ANALYSIS OF JUDICIAL DECISIONS: JUDICIAL DISRUPTION AND PRACTICES
Lluc Font Pomarol

D

List of top-novelty judicial decisions

- Sentencia de la Audiencia Provincial de Madrid de 14 de julio de 2017 Num. 510/2017
- Sentencia de la Audiencia Provincial de Madrid de 10 de noviembre de 2017 Num. 730/2017
- Sentencia de la Audiencia Provincial de Madrid de 9 de junio de 2017 Num. 402/2017
- Sentencia de la Audiencia Provincial de Madrid de 12 de mayo de 2017 Num. 321/2017
- Sentencia de la Audiencia Provincial de Madrid de 8 de enero de 2017 Num. 1/2018
- Sentencia de la Audiencia Provincial de Madrid de 22 de octubre de 2018 Num. 700/2018
- Sentencia de la Audiencia Provincial de les Illes Balears de 21 de febrero de 2018 Num. 29/2018
- Sentencia de la Audiencia Provincial de Madrid de 4 de septiembre de 2018 Num. 581/2018
- Sentencia de la Audiencia Provincial de Albacete de 30 de octubre de 2018 Num. 428/2018
- Sentencia de la Audiencia Provincial de Navarra de 30 de junio de 2016 Num. 348/2016
- Sentencia de la Audiencia Provincial de La Coruña de 11 de diciembre de 2018 Num. 532/2017
- Auto del Tribunal Supremo de 17 de febrero de 2016
- Auto del Tribunal Supremo de 13 de diciembre de 2017
- Auto del Tribunal Supremo de 17 de mayo de 2017
- Sentencia de la Audiencia Provincial de Málaga de 6 de noviembre de 2003 Num. 574/2003

- Sentencia de la Audiencia Provincial de Málaga de 21 de julio de 2002 Num. 394/2002
- Auto de la Audiencia Provincial de Cádiz de 27 de Febrero de 2003 Num. 64/2011
- Auto de la Audiencia Provincial de Madrid de 4 de septiembre de 2012 Num. 682/2012
- Sentencia de la Audiencia Provincial de Madrid de 15 de marzo de 2011 Num. 170/2011
- Sentencia del Tribunal Supremo de 13 de octubre de 2015 Num. 558/2015
- Auto de la Audiencia Provincial de Castellón de 21 de julio de 2014 Num. 164/2014
- Sentencia de la Audiencia Provincial de Madrid de 4 de septiembre de 2013 Num. 395/2013
- Sentencia de la Audiencia Provincial de Sevilla de 20 de marzo de 2013 Num. 135/2013
- Sentencia de la Audiencia Provincial de Valladolid de 25 de septiembre de 2013 Num. 344/2013
- Sentencia de la Audiencia Provincial de Madrid de 3 de mayo de 2017 Num. 267/2017

Bibliography

Ackoff, Russell (1989). From data to wisdom. In: *Journal of Applied Systems Analysis*. Vol. 16, pp. 3–9.

Agteren, Joep van et al. (2021). A systematic review and meta-analysis of psychological interventions to improve mental wellbeing. *Nature Human Behaviour* 5.5, pp. 631–652.

Ainsworth, Janet and Patrick Juola (2018). Who wrote this?: Modern forensic authorship analysis as a model for valid forensic science. *Washington University Law Review* 96, pp. 1161–1189.

Alexander, Charlotte S. and Mohammad Javad Feizollahi (2020). On dragons, caves, teeth, and claws: legal analytics and the problem of court data access. *Computational Legal Studies*, pp. 95–123.

Alschner, Wolfgang (2020). Sense and similarity: automating legal text comparison. *Computational Legal Studies: The Promise and Challenge of Data-Driven Research*, pp. 9–28.

Andrei, Victor and Ognjen Arandjelović (2016). Complex temporal topic evolution modelling using the Kullback-Leibler divergence and the Bhattacharyya distance. *EURASIP J Bioinform Syst Biol.* 16.

Argamon, Shlomo, Moshe Koppel, James W. Pennebaker, and Jonathan Schler (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM* 52.2, pp. 119–123.

Ash, Elliott et al. (2022). *Gender Attitudes in the Judiciary: Evidence from U.S. Circuit Courts*. Tech. rep.

Asmat, Roberto and Lajos Kossuth (2021). Gender Differences in Judicial Decisions under Incomplete Information: Evidence from Child Support Cases. *SSRN Electronic Journal*, pp. 1–39.

Bagues, Manuel, Mauro Sylos-Labini, and Natalia Zinovyeva (2017). Does the gender composition of scientific committees matter? *American Economic Review* 107.4, pp. 1207–1238.

Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18.2, pp. 135–160.

Barabási, Albert László and Réka Albert (1999). Emergence of Scaling in Random Networks. *Science* 286.5439, pp. 509–512.

Barron, Alexander T. J., Jenny Huang, Rebecca L. Spang, and Simon DeDeo (2018). Individuals, institutions, and innovation in the debates of the French Revolution. *Proceedings of the National Academy of Sciences*.

Baude, W., A.S. Chilton, and Anup Malani (2017). Making doctrinal work more rigorous: Lessons from systematic reviews. *University of Chicago Law Review* 84, pp. 37–58.

Bertrand, Marianne and Sendhil Mullainathan (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *Inequality in the 21st Century: A Reader* 1996, pp. 304–308.

Blei, David, Lawrence Carin, and David Dunson (2010). Probabilistic Topic Models. *IEEE Signal Processing Magazine* 27.6, pp. 55–65.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3.null, 993–1022.

Bollobás, Béla (2001). Random Graphs. 2nd ed. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Boulesteix, Anne-Laure, Silke Janitza, Jochen Kruppa, and Inke R. König (2012). Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.6, pp. 323–329.

Boyd, Christina L., Lee Epstein, and Andrew D. Martin (2010). Untangling the causal effects of sex on judging. *American Journal of Political Science* 54.2, pp. 389–411.

Breiman, Leo (2001). Random forests. *Machine Learning* 45, pp. 5–32.

Cafiero, Florian and Jean Baptiste Camps (2019). Why Molière most likely did write his plays. *Science Advances* 5.11.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan (2017). *Semantics derived automatically from language corpora contain human-like biases*. Tech. rep., pp. 183–186.

Charlotin, Damien (2020). ""Authorities" in International Dispute Settlement: a Data Analysis (Doctoral thesis)". PhD thesis. University of Cambridge.

Christensen, Martin Lolle, Henrik Palmer Olsen, and Fabian Tarissan (2016). Identification of Case Content with Quantitative Network Analysis: An Example from the ECtHR. English. In: *Legal Knowledge and Information Systems*. Ed. by Floris Bex and Serena Villata. null ; Conference date: 14-12-2016 Through 16-12-2016. United States: IOS Press, pp. 53–62.

Citron, Daniel T. and Paul Ginsparg (2015). Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences* 112.1, pp. 25–30.

Cohen, Alma and Crystal S. Yang (2019). Judicial politics and sentencing decisions. *American Economic Journal: Economic Policy* 11.1, pp. 160–191.

Collins, Paul M., Kenneth L. Manning, and Robert A. Carp (2010). Gender, critical mass, and judicial decision making. *Law and Policy* 32.2, pp. 260–281.

Crow, Matthew S. and Natalie Goulette (2022). Judicial diversity and sentencing disparity across U.S. District Courts. *Journal of Criminal Justice* 82.May, p. 101973.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso (2011). Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences* 108.17, pp. 6889–6892.

De Lucio, Juan and Juan S. Mora-Sanguinetti (2021). New Dimensions of Regulatory Complexity and Their Economic Cost. An Analysis Using Text Mining. *SSRN Electronic Journal*.

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, pp. 257–277.

Devine, Patricia G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology* 56.1, pp. 5–18.

Domènech-Pascual, Gabriel (2021). Thought Experiments in Law. *Law and Method*, pp. 1–21.

Domènech-Pascual, Gabriel (2022). Nacimiento, consolidación y persistencia de las teorías jurídicas defectuosas, pp. 1–46.

Dyevre, Arthur (2021). The promise and pitfall of automated text-scaling techniques for the analysis of jurisprudential change. *Artificial Intelligence and Law* 29.2, pp. 239–269.

Eck, Kristine and Charles Crabtree (2020). Gender differences in the prosecution of police assault: Evidence from a natural experiment in Sweden. *PLoS ONE* 15.7 July.

Erdős, Paul and Alfréd Rényi (1959). On Random Graphs. *Publicationes Mathematicae* 6, pp. 290–297.

Evans, James A. and Pedro Aceves (2016). Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology* 42.1, pp. 21–50.

Evans, Michael C., Wayne V. McIntosh, Jimmy Lin, and Cynthia L. Cates (2011). Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research. *SSRN Electronic Journal*.

Fontana, Magda (2020). New and atypical combinations : An assessment of novelty and interdisciplinarity. 49.November 2018.

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, pp. 324–346.

Friedrich, Roland (2021). Complexity and Entropy in Legal Language. *Frontiers in Physics* 9.June, pp. 1–11.

Garcia-Teruel, Rosa Maria and Sergio Nasarre-Aznar (2022). Quince años sin solución para la vivienda. La innovación legal y la ciencia de datos en política de vivienda. *Revista Crítica de Derecho Inmobiliario*.

García-Gavilanes, Ruth, Anders Mollgaard, Milena Tsvetkova, and Taha Yasseri (2017). The memory remains: Understanding collective memory in the digital age. *Science Advances* 3.4, e1602368.

Garg, Amit X., Dan Hackam, and Marcello Tonelli (2008). Systematic review and meta-analysis: When one study is just not enough. *Clinical Journal of the American Society of Nephrology* 3.1, pp. 253–260.

Gerlach, Martin, Tiago P. Peixoto, and Eduardo G. Altmann (2018). A network approach to topic models. *Science Advances* 4.7.

Gerlach, Martin, Hanyu Shi, and Luís A. Nunes Amaral (2019). A universal information theoretic approach to the identification of stopwords. *Nature Machine Intelligence* 1.12, pp. 606–612.

Gestel, Rob van and Hans-Wolfgang Micklitz (2014). Why Methods Matter in European Legal Scholarship. *European Law Journal* 20.3, pp. 292–316.

Greenwald, Anthony G, Debbie E Mcghee, and Jordan L K Schwartz (1998). *Measuring Individual Differences in Implicit Cognition: The Implicit Association Test*. Tech. rep. 6, pp. 1464–1480.

Grimmer, Justin and Brandon M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21.3, pp. 267–297.

Guimerà, Roger and Marta Sales-Pardo (2011). Justice Blocks and Predictability of U.S. Supreme Court Votes. *PLOS ONE* 6.11, pp. 1–8.

Hall, Mark A. and Ronald F. Wright (2008). Systematic Content Analysis of Judicial Opinions. *California Law Review* 96.1, pp. 63–122.

Harris, Allison P. and Maya Sen (2019). Bias and judging. *Annual Review of Political Science* 22, pp. 241–259.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2001). The Elements of Statistical Learning. Data Mining, Inference, and Prediction. 2nd ed. Springer Series in Statistics. Springer New York.

Hesselink, Martijn (2009). A European legal method? On European private law and scientific method. *European Law Journal* 15.1, pp. 20–45.

Hillyard, Paddy (2007). Law's Empire: Socio-legal Empirical Research in the Twenty-first Century. 34.2, pp. 266–279.

Hosseini, Monireh and Zohreh Tammimy (2016). Recognizing users gender in social media using linguistic features. *Computers in Human Behavior* 56, pp. 192–197.

Hughes, James M., Nicholas J. Foti, David C. Krakauer, and Daniel N. Rockmore (2012). Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences* 109.20, pp. 7682–7686.

Hutchinson, Terry and Nigel Duncan (2012). Defining and Describing What We Do: Doctrinal Legal Research. *Deakin Law Review* 17, pp. 83–119.

Iacopini, Iacopo, Staša Milojević, and Vito Latora (2018). Network Dynamics of Innovation Processes. *Physical Review Letters* 120.4, pp. 1–6.

Itti, Laurent and Pierre Baldi (2009). Bayesian surprise attracts human attention. *Vision Research* 49.10, pp. 1295–1306.

Jaynes, Edwin T. (2003). Probability theory: The logic of science. Cambridge university press.

Jockers, Matthew L. and David Mimno (2013). Significant themes in 19th-century literature. *Poetics* 41.6, pp. 750–769.

Juola, Patrick (2008). Authorship Attribution. *Foundations and Trends in Information Retrieval* 1.3, pp. 233–334.

Katz, Daniel, Corinna Coupette, Janis Beckedorf, and Dirk Hartung (2020). Complex Societies and the Growth of the Law. *Scientific Reports* 10, p. 18737.

Katz, Daniel Martin and M. J. Bommarito (2014). Measuring the complexity of the law: the United States Code. Vol. 22. 4, pp. 337–374.

Kestemont, Mike (2014). Function Words in Authorship Attribution From Black Magic to Theory? *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*, pp. 59–66.

Klingenstein, Sara, Tim Hitchcock, and Simon DeDeo (2014). The civilizing process in London's Old Bailey. *Proceedings of the National Academy of Sciences*.

Koning, Rembrand, Sampsa Samila, and John Paul Ferguson (2021). Who do we invent for? Patents by women focus more on women's health, but few women get to invent. *Science* 372.6548, pp. 1345–1348.

Kort, Fred (1957). Predicting Supreme Court Decisions Mathematically: A Quantitative Analysis of the "Right to Counsel" Cases. *The American Political Science Review* 51.1, pp. 1–12.

Krenn, Mario and Antonc Zeilinger (2020). Predicting research trends with semantic and neural networks with an application in quantum physics. *Proceedings of the National Academy of Sciences of the United States of America* 117.4, pp. 1910–1916.

Kuhn, Tobias, Matja ž Perc, and Dirk Helbing (2014). Inheritance Patterns in Citation Networks Reveal Scientific Memes. *Phys. Rev. X* 4 (4), p. 041036.

Kulik, Carol T., Elissa L. Perry, and Molly B. Pepper (2003). Here comes the judge: The influence of judge personal characteristics on federal sexual harassment case outcomes. *Law and Human Behavior* 27.1, pp. 69–86.

Kullback, S. and R. A. Leibler (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics* 22.1, pp. 79–86.

Lancichinetti, Andrea et al. (2015). High-Reproducibility and High-Accuracy Method for Automated Topic Classification. *Phys. Rev. X* 5 (1), p. 011007.

Landes, William M., Lawrence Lessig, and Michael E. Solimine (1998). Judicial influence: A citation analysis of federal courts of appeals judges. *Journal of Legal Studies* 27.2 PART I, p. 271.

Langford, Malcolm, Daniel Behn, and Runar Lie (2020). Computational stylometry: predicting the authorship of investment arbitration awards. *Computational Legal Studies: The Promise and Challenge of Data-Driven Research*, pp. 53–76.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin (2013). Bias in Peer Review. *Journal of the American Society for Information Science and Technology* 64.July, pp. 1852–1863.

Levina, Anna, Viola Priesemann, and Johannes Zierenberg (2022). Tackling the subsampling problem to infer collective properties from limited data. *Nature Reviews Physics*, pp. 1–15.

Livermore, Michael A, Allen B Riddell, and N Daniel (2016). The Supreme Court and the Judicial Genre.

Lupu, Yonatan and James H. Fowler (2013). Strategic citations to precedent on the U.S. Supreme Court. *Journal of Legal Studies* 42.1, pp. 151–186.

Lupu, Yonatan and Erik Voeten (2011). Precedent in international courts: A network analysis of case citations by the European court of human rights. *British Journal of Political Science* 42.2, pp. 413–439.

Manning, Christorpher D., Prabhakar Raghavan, and Hinrich Schütze (2008). An introduction to information retrival. Cambridge Univeristy Press.

Medvedeva, Masha, Michel Vols, and Martijn Wieling (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law* 28, pp. 237–266.

Mones, Enys, Piotr Sapieżyński, Simon Thordal, Henrik Palmer Olsen, and Sune Lehmann (2021). Emergence of network effects and predictability in the judicial system. *Scientific Reports* 11.1, pp. 1–10.

Moss-Racusin, Corinne A., John F. Dovidio, Victoria L. Brescoll, Mark J. Graham, and Jo Handelsman (2012). Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences of the United States of America* 109.41, pp. 16474–16479.

Murdock, Jaimie, Colin Allen, and Simon DeDeo (2017). Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks. *Cognition* 159, pp. 117–126.

Nasarre-Aznar, Sergio (2020). Los años de la crisis de la vivienda. De las hipotecas subprime a la vivienda colaborativa. Valencia: Tirant lo Blanch.

Nasarre-Aznar, Sergio and Rosa Maria Garcia-Teruel (2018). Evictions and homelessness in Spain 2010–2017. In: *Loss of Homes and Evictions across Europe*. Ed. by Padraic Kenna, Sergio Nasarre-Aznar, Peter Sparkes, and Christoph U. Schmid. Massachusetts 01060 USA: Edward Elgar Publishing, 292–332.

Neidorf, Leonard, Madison S. Krieger, Michelle Yakubek, Pramit Chaudhuri, and Joseph P. Dexter (2019). Large-scale quantitative profiling of the Old English verse tradition. *Nature Human Behaviour* 3.6, pp. 560–567.

Nemenman, Ilya, William Bialek, and Rob De Ruyter Van Steveninck (2004). Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E* 69.5, p. 056111.

Nemenman, Ilya, Fariel Shafee, and William Bialek (2001). Entropy and inference, revisited. *Advances in neural information processing systems* 14.

Newman, M. E.J. (2003). The structure and function of complex networks. *Society for Industrial and Applied Mathematics Review* 45.2, pp. 167–256.

Newman, Matthew L., Carla J. Groom, Lori D. Handelman, and James W. Pennebaker (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45.3, pp. 211–236.

Olsen, Henrik Palmer and Aysel Küçüksu (2017). Finding hidden patterns in ECtHR's case law: On how citation network analysis can improve our knowledge of ECtHR's Article 14 practice. *International Journal of Discrimination and the Law* 17.1, pp. 4–22.

Pah, Adam R. et al. (2020). How to build a more open justice system. *Science* 369.6500, pp. 134–136.

Paley, Andrew et al. (2021). From Data to Information : Automating Data Science to Explore. *ICAIL '21: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pp. 119–128.

Panagis, Yannis, Martin Christensen, and Urška Šadl (2016). On top of topics: leveraging topic modeling to study the dynamic case-law of international courts of Law, iCourts centre of excellence for international courts. *Legal Knowledge and Information Systems* 294, pp. 161 –166.

Peixoto, Tiago P. (2014). Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X* 4.1, pp. 1–18.

Peixoto, Tiago P. (2019). Bayesian Stochastic Blockmodeling. In: *Advances in Network Clustering and Blockmodeling*. John Wiley & Sons, Ltd. Chap. 11, pp. 289–332.

Piga, Angelo, Lluc Font-Pomarol, Marta Sales-Pardo, and Roger Guimerà (2023). Bayesian estimation of information-theoretic metrics for sparsely sampled distributions.

Posner, Richard A. (1995). Judges' Writing Styles (And Do They Matter?) *The University of Chicago Law Review* 62.4, p. 1421.

Posner, Richard A. (2000). An Economic Analysis of the Use of Citations in the Law. *American Law and Economics Review* 2.2, pp. 381–406.

Quemy, Alexandre and Robert Wrembel (2020). On Integrating and Classifying Legal Text Documents. In: *Database and Expert Systems Applications*. Ed.

by Sven Hartmann, Josef Küng, Gabriele Kotsis, A. Min Tjoa, and Ismail Khalil. Cham: Springer International Publishing, pp. 385–399.

Rice, Douglas, Jesse H. Rhodes, and Tatishe Nteta (2019). Racial bias in legal language. *Research and Politics* 6.2.

Rissanen, Jorma (1978). Modelling by Shortest Data Description. *Automatica* 14, pp. 465–471.

Rockmore, Daniel N., Chen Fang, Nicholas J. Foti, Tom Ginsburg, and David C. Krakauer (2018). The cultural evolution of national constitutions. *Journal of the Association for Information Science and Technology* 69.3, pp. 483–494.

Rowley, Jennifer (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science* 33.2, pp. 163–180.

Rudin, Cynthia (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1.5, pp. 206–215.

Rutherford, Alex et al. (2018). Inferring mechanisms for global constitutional progress. *Nature Human Behaviour* 2, pp. 592–599.

Rzhetsky, Andrey, Jacob G Foster, Ian T Foster, and James A Evans (2015). Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences of the United States of America* 112.47, pp. 14569–74.

Šadl, Urška and Henrik Palmer Olsen (2017). Can Quantitative Methods Complement Doctrinal Legal Studies? Using Citation Network and Corpus Linguistic Analysis to Understand International Courts. *Leiden Journal of International Law* 30.2, pp. 327–349.

Savoy, Jacques (2013). Authorship attribution based on a probabilistic topic model. *Information Processing and Management* 49.1, pp. 341–354.

Shaffer, Gregory and Tom Ginsburg (2012). The empirical turn in international legal scholarship. *The Americal Journal of International Law* 106.1, pp. 1–46.

Sheshadri, Karthik and Munindar P. Singh (2019). The public and legislative impact of hyperconcentrated topic news. *Science Advances* 5.8, eaat8296.

Shlens, Jonathon (2014). Notes on Kullback-Leibler Divergence and Likelihood. *CoRR* abs/1404.2000.

Stoeger, Thomas, Martin Gerlach, Richard I. Morimoto, and Luís A. Nunes Amaral (2018). Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biology* 16.9, pp. 1–25.

Stolker, C J J M (2005). Legal Journals: in Pursuit of a More Scientific Approach. *European Journal of Legal Education* 2.2, pp. 77–94.

Tarissan, Fabien and Raphaëlle Nollez-Goldbach (2016). Analysing the first case of the International Criminal Court from a network-science perspective. *Journal of Complex Networks* 4.4, p. 616.

Uzzi, Brian, Satyam Mukherjee, Michael Stringer, and Ben Jones (2013). Atypical Combinations and Scientific Impact. *Science* 342.6157, pp. 466–468.

Van Kuppevelt, Dafne, Gijs Van Dijck, and Marcel Schaper (2020). Purposes and challenges of legal citation network analysis on case law. *Computational Legal Studies: The Promise and Challenge of Data-Driven Research* 2020, pp. 265–292.

Venzke, Ingo (2015). International law and its methodology: Introducing a new leiden Journal of International Law series. *Leiden Journal of International Law* 28.2, pp. 185–187.

Vick, Douglas W. (2004). Interdisciplinarity and the Discipline of Law. *Journal of Law and Society* 31.2, pp. 163–193.

Watts, Duncan (2007). A twenty-first century science. *Nature* 445, p. 489.

Watts, Duncan J and Steven H Strogatz (1998). Collective dynamics of 'small-world' networks. *Nature* 393, pp. 440–442.

Welch, Susan, Michael Combs, and John Gruhl (1988). Do Black Judges Make a Difference? *American Journal of Political Science* 32.1, pp. 126–136.

Whalen, Ryan (2016). Legal Networks: The promises and the challenges of legal network analysis. *Michigan State Law Review* 539.

Wolpert, David H. and David R. Wolf (1995). Estimating functions of probability distributions from a finite set of samples. *Physical Review E* 52.6, p. 6841.

Wu, Lingfei, Dashun Wang, and James A. Evans (2019). Large teams develop and small teams disrupt science and technology. *Nature* 566.7744, p. 378.