2017

Mass Digitization of Chinese Court Decisions: How to Use Text as Data in the Field of Chinese Law

Benjamin L. Liebman
Columbia Law School, bliebm@law.columbia.edu

Margaret Roberts
meroberts@ucsd.edu

Rachel E. Stern
restern@fas.harvard.edu

Alice Wang
alice.z.wang@runbox.com

UC San Diego
SCHOOL OF GLOBAL POLICY AND STRATEGY

21ST CENTURY CHINA CENTER

Mass Digitization of Chinese Court Decisions: How to Use Text as Data in the Field of Chinese Law

21st Century China Center Research Paper No. 2017-01

Margaret E. Roberts

Benjamin L. Liebman

Rachel E. Stern

Alice Z. Wang

Mass Digitization of Chinese Court Decisions: How to Use Text as Data in the Field of Chinese Law *

Benjamin L. Liebman[†] Margaret E. Roberts[‡] Rachel E. Stern[§]
Alice Z. Wang [¶]

June 2017

Abstract

Over the past five years, Chinese courts have placed tens of millions of court judgments online. We analyze the promise and pitfalls of using this remarkable new data source through the construction and examination of a dataset of 1,058,990 documents from Henan province. Courts posted judgments in roughly half of all cases in 2014 and, although the percent of cases posted online has likely risen since then, the single greatest challenge facing researchers remains documenting gaps in the data. We find that missing data varies widely by court, and that intermediate courts disclose significantly more documents than basic level courts. But court level, GDP per capita, population, and mediation rates are insufficient fully to explain variation in disclosure rates. Further work is needed to better understand how resources and incentives might be skewing the data. Despite incomplete information, however, a topic model of 20,321 administrative court judgments demonstrates how mass digitization of court decisions opens a new window into the practice of everyday law in China. Unsupervised machine learning combined with close reading of selected cases reveals surprising trends in administrative disputes as well as important research questions. Taken together, our findings suggest a need for humility and methodological pluralism among scholars seeking to use large-scale data from Chinese courts. The vast amount of incomplete data now available may frustrate attempts to find quick answers to existing questions, but the data excel at opening new pathways for research and at adding nuance to existing assumptions about the role of courts in Chinese society.

*All authors contributed equally to this project.

[†]Robert L. Lieff Professor of Law and Director of the Chinese Legal Studies, Columbia Law School, bliebm@law.columbia.edu.

[‡]Assistant Professor, Department of Political Science, University of California, San Diego, meroberts@ucsd.edu.

[§]Assistant Professor of Law and Political Science, University of California, Berkeley, rstern@law.berkeley.edu.

[¶]Postdoctoral Fellow, Center for Chinese Legal Studies, Columbia Law School, alice.z.wang@runbox.com.

1 Introduction

In January 2014, every level of Chinese courts became responsible for uploading judicial decisions to a centralized website run by the Supreme People's Court (SPC) (SPC 2013). This policy, which followed and formalized years of local efforts to put cases online, has led to a rapid expansion of the public record of court activity: more than twenty-nine million documents were posted by June 2017.[1] Although the resulting database is far from complete, this new source offers an unusual opportunity to transform our understanding of the Chinese legal system by developing a granular portrait of what happens in courts every day.

The release of tranches of judicial documents also coincides with growing interest across the social sciences in treating text as data and using computer science tools to uncover patterns in document collections too big for any research team to code by hand.[2] As others have pointed out (Lucas et al. 2015), there is a natural fit between automated content analysis and the comparative study of law, politics and society. Countries around the world are producing public text at unprecedented rates, from social media posts to political press releases, and there is much to be learned in digesting it. Computers can help systematically map the "great unread" by analyzing patterns of topic prevalence, word use, and tone inside a corpus (Miller 2013). And new techniques of reading at a distance can also be paired with the close reading that has long been a staple of scholarship. A small number of scholars have begun to apply such techniques to data gleaned from authoritarian regimes, including China.

This article stands at the intersection of these two trends, toward the digitization and public release of court records in China on the one hand and computational social science on the other. It is meant as a field guide to the promise and pitfalls of using large numbers of court decisions to track China's legal and political evolution and, more generally, of using text as data to study Chinese institutions. Our motivation is a shared sense that this new source is impossible to ignore, particularly as political conditions make interviews and surveys increasingly difficult, and can teach us a great deal. Yet the arrival of big data to the field of Chinese law is also an intimidating prospect, particularly because existing databases are imperfect. Researchers must grapple with a long list of challenges: holes in the public record and little information about what is missing, frequently changing rules on what courts should post, inconsistent formatting of court documents, the in-flux nature of databases that routinely expand as new documents come online but also sometimes remove older material, and the technological difficulty of large-scale downloads from websites not designed for that purpose.

In addition, court decisions themselves are an incomplete source. Courts avoid politically sensitive cases and their decisions are often terse summations of facts, claims and outcomes. Much is missing, in particular any record of behind-the-scenes interaction between courts, government agencies and the Communist Party. Nevertheless, the last decade of scholar-

1. The SPC website provides a real-time calculation of the total number of documents. See http://wenshu.court.gov.cn/.

2. For good overviews of how social scientists have been using text as data, see (Grimmer and Stewart 2013), (Lucas et al. 2015), and (Evans and Aceves 2016).

ship shows growing recognition of how judicial decisions can open a window onto topics as diverse as corporate law (Howson 2010), medical malpractice litigation (Liebman 2015), tax administration (Li 2013; Cui, forthcoming), environmental crime (Stern 2014), judicial decision-making (Stern 2013; He and Su 2013), changing labor relations (Huang 2017) and criminal sentencing (Liebman 2015). Some of these efforts, especially earlier on, were based on documents obtained through personal contacts. In recent years, however, scholars have also begun to download cases manually from various websites to create, read, and sometimes code a collection of cases on a particular topic. The next shift is relying on computer assistance to download and analyze larger bodies of text, but working at this scale presents an array of technological and methodological challenges.

Anticipating this shift in scholarship, our goals in this article are descriptive (what is available?), prescriptive (what is the best way to use large scale databases like this one?), and theoretical (how will the availability of big data affect the study of Chinese institutions?). The article opens with a short history of the dramatic expansion of public access to court documents. We then turn to a hands-on illustration of the promise and pitfalls of a large-scale computational approach, based on a dataset of 1,058,990 documents downloaded from a website set up by the Henan High People's Court.[3] The single biggest challenge is documenting bias in the data, as many cases are missing from the online record. We find wide intra-provincial variation in judicial disclosure rates, which persists even after accounting for court level, mediation rates, GDP per capita and population. This suggests it is risky to assume that any case database—no matter how large—accurately reflects court dockets. At the same time, however, we demonstrate how treating text as data can yield important insights even with data we know to be incomplete. We use a topic model of 20,321 administrative court judgments from Henan to show how combining unsupervised machine learning with close reading of selected cases can unearth both surprising trends in Chinese court practice and important research questions.

2 The Origins of Judicial Disclosure in China

As a style of governance, transparency is associated with democracies because sharing information seems intuitively linked to an aspiration for political participation widespread enough to oblige responsive, accountable government. Even though few authoritarian states show much interest in empowering citizens this way, transparency has entered the Chinese lexicon of governance over the past decade. Freedom-of-information regulations passed in 2007 (Horsley 2007, 2010), co-exist with a media spotlight glaring enough to expose at least some malfeasance (Liebman 2005). Officials are also required to release certain types of data, including selected environmental statistics (Lorentzen, Landry, and Yasuda 2013) and, now, court decisions.

3. Our dataset covers all documents on the site as of November 29, 2015. After this date, Henan courts started uploading new cases exclusively to the SPC website. Future work will compare the content of the Henan provincial website to Henan court documents made available contemporaneously and subsequently on the SPC website.

Yet even given the embrace of modest forms of "controlled transparency" (Liebman 2011), there has been a remarkable shift in China's courts in recent years. As recently as five years ago, it was difficult for scholars to construct even small datasets of court opinions without personal contacts in the courts. Today, more than 29 million cases are online - a number China's SPC claims makes its website the largest collection of public cases in the world (Sina 2016).[4] Certainly, China's commitment to making court opinions available online is unusual - if not unique - in the authoritarian world and not common even among democratic civil law systems.[5]

As with any new corpus, the intelligent use of case databases requires understanding how they came to exist. Most court judgments in China have technically long been public documents, meaning that a Chinese citizen has the right to view a case at the courthouse.[6] In practice, however, for most of the past three decades court decisions were typically available only to the people directly involved in the case.

In the early years of the reform era, the SPC recognized the value of making some court decisions publicly available as a means of educating both judges and litigants. In 1985, the SPC began publishing its official gazette (最高人民法院公报), which included a small number of selected cases. Although not formally recognized as precedent, published cases were meant to guide lower courts on how to handle particular points of law. The idea was that model cases could boost access to legal information and, in so doing, also improve the uniformity and quality of court decisions across China. Other publications followed suit, such as the People's Court News, which started including court opinions alongside news items about or of interest to the judiciary soon after it was established in 1992. Collections of representative cases, some curated by the SPC and others by academics or local courts, also started to become standard fare for legal publishers and a source of practical guidance for their readers (Liebman and Wu 2007, 289).

In the late 1990s and early 2000s, individual courts began posting selected "representative cases" (典型案例) online. How much material was available varied greatly, from a handful of decisions to hundreds of cases, with a few standouts attempting to cultivate a reputation for innovation by pursuing transparency. Fee-based databases also started to sprout up, as new actors saw a commercial opportunity to build the Chinese version of America's Lexis-Nexis and Westlaw. These early commercial entrants aggregated national and local laws and regulations, which were previously difficult to find in one place, and also provided access to

4. The claim is difficult to verify, given the wide range of practices in other countries, the role of private companies in providing such services in the U.S. and elsewhere, and the fact that many jurisdictions leave the posting of court decisions to individual courts.

5. There is a small political science literature on the purposes and effects of authoritarian transparency (Malesky, Schuler, and Tran 2012; Lorentzen, Landry, and Yasuda 2013; Hollyer, Rosendorff, and Vreeland 2015). Transparency is a capacious word, however, and we are not aware of any work focused on court transparency in authoritarian regimes

6. Article 156 of the Civil Procedure Law discusses the public's right to refer to final court decisions (查阅), except for those involving state secrets, trade secrets, or relating to personal privacy (NPC 2013). In 2015, the SPC further specified that members of the public can request access to court decisions by submitting a written request with the case number or the name of the parties.(SPC 2015a)

selected court documents. The number of documents in these commercial databases grew quickly as they competed with each other to build the biggest and most complete database, although coverage remained spotty.[7] A few law firms also started posting examples of cases handled by their lawyers, presumably to project professionalism and attract clients.

In the early 2000s, liberal scholars began to call for courts to place all opinions online. Writing in *Southern Weekend*, for example, Peking University law professor He Weifang called for the end of the era of "closed justice" (封闭司法) and predicted transparency would yield manifold benefits, from reduced corruption to restored public confidence in courts (W. He 2003). Somewhat ironically, it was not until Wang Shengjun became president of the SPC in 2008 that the SPC itself made a push to place large numbers of court decisions online. Wang, who lacked legal training prior to becoming China's chief judge, was widely perceived to be ideologically conservative. Still, under his leadership, the SPC endorsed judicial transparency as a goal in its third five-year plan for legal reform in 2009 and also publicly encouraged lower court efforts to compile and publish judicial decisions (SPC 2009a, 2009b).[8]

These cues prompted a raft of local initiatives, including in Henan, where a mid-2009 Henan High Court order mandated that all courts in the province place the vast majority of decisions online.[9] This new requirement followed a wave of high profile wrongful convictions in Henan, and was one of a number of populist measures adopted by Henan High Court President Zhang Liyong (Liebman 2015, 161-162). Other provinces followed Henan's lead, with provincial court websites hosting tens and even hundreds of thousands of court judgments.

By the time the SPC called on courts nationwide to begin posting most cases online in 2013 (SPC 2013),[10] this policy was less of an about-face than a studied choice to build on local experiments nationwide.[11] The new rules created a centralized website, called "China Court Judgments"(中国裁判文书网), which launched on July 1, 2013. Although the website initially only hosted SPC decisions, the website included documents from across the country by the middle of 2015 (Sina 2016). In Henan, a period of overlap when courts posted cases to both the Henan High Court website and the SPC website ended in late November 2015

7. None of these databases publicly disclosed how they obtained court documents, or how many similar cases were not publicly available. In the mid-2000s, employees at one of the leading commercial databases said they obtained cases by collecting publicly available cases from the internet and case compendia, and also through "cooperative relationships" (合作关系) with individual courts see (Stern 2013, 124).

8. The SPC's regulation was permissive, not mandatory. It stated that "the people's courts may, according to the needs of legal advocacy, law research, case guidance and unification of standards for judgment, compile, print and publish various judgment documents in a centralized way." See (SPC 2009b).

9. Initially, none of this material was meant to be permanently available: the Henan regulations stated that cases should be posted for one year only. This policy changed in 2012, when the Henan High People's Court launched its own website that aggregated cases from all courts in the province and stopped removing cases that were more than a year old.

10. An earlier SPC notice, issued in 2010, had taken a permissive approach, stating that courts could post cases online, subject to certain exceptions. For a discussion of how judges, lawyers, and scholars reacted to the new policy, see (Liebman 2015, 163-164).

11. For more on the Chinese tradition of giving local officials leeway to try out new approaches before adopting good ideas nationwide, see (Heilmann and Perry 2011).

when the provincial website started redirecting visitors to the SPC site.

Alongside this national policy change, new commercial players entered the market for legal information. Upstarts such as Wusong in Beijing and OpenLaw in Shanghai, both founded in 2014, arose to challenge established companies, such as Peking University's Beida Fabao and Beida Fayi. Today, websites run by commercial operators often contain more documents than the official SPC website and provide additional functionality.[12]

Overall, the SPC's approach to expanding public access has been to move gradually toward increasing the volume and scope of the material available, but to grant lower courts discretion over how constraints are interpreted and implemented. The first set of SPC rules governing the public release of court opinions, issued in 2013, provided exemptions for certain types of cases: cases involving state secrets or personal privacy, juvenile criminal cases, disputes concluded through mediation, and other documents deemed "inappropriate" (不宜) to publicize.[13] Likewise, the SPC initially did not require either enforcement decisions or official notification of withdrawals to be posted online. These restrictions are significant: the volume of cases resolved through mediation is large, and many of the most controversial cases in China are either mediated or could be considered inappropriate for publication.

Over time, however, the SPC has unmistakably pushed lower courts to make more information publicly available. A second round of SPC rules clarifying what courts should post, issued as a judicial interpretation in 2016, expanded the public record in important ways (SPC 2016a). In addition to criminal, civil, and administrative decisions, courts now must post a range of documents only sporadically made public in the past. These include outcomes in state compensation proceedings, changes in criminal sentences, mediated administrative cases, enforcement decisions and withdrawals. The general principle is that any document that reflects termination of a case should be made public unless it falls into a specific excluded category. Most important, courts are also required to release the case number of any decision deemed unsuitable for posting online, and explain why the judgment

12. This is in part because the sites go further back in time, including some cases from prior to the launch of the SPC site. Yet even for the time period covered by the SPC site, Openlaw and Itslaw have more cases than the official SPC site. The reason for this is unclear, but is likely because companies sometime add cases scraped from provincial websites or obtained directly from parties and lawyers to those available on the SPC website. Commercial websites may also include duplicate copies of the same case, one downloaded from a provincial court website and another from the SPC website. In general, regulatory uncertainty continues to cloud this market. A notice on the SPC website forbids the commercial establishment of a "mirror website," and prohibits using information provided on the website to obtain "illegal benefit." Although the legal basis of these restrictions on commercial use is unclear, commercial websites have deployed a number of strategies to avoid falling afoul of the SPC. Openlaw, for example, is registered as a non-profit and does not sell any information on its website. Likewise, Wusong's case database is open access. Instead of charging for access to laws, regulations and cases, the company is bridging into the market for legal services through a fee-based service that matches clients with lawyers who meet their needs and budget. Wusong also provides litigation consulting to companies, examining prior cases for insights on issues such as where to file a case and the likelihood of success in litigation.

13. The rules also called for certain names to be redacted, including parties in divorce and inheritance cases, witnesses in criminal cases, crime victims, and certain first time defendants in minor criminal cases. The rules also called for some personal information to be redacted.

was held back.[14] If followed, the new practice would provide a rough indicator of the percent of the docket placed online, and offer insight into the reasons cases are not made public.

However, the 2016 rules reiterate the principle that local courts have discretion not to post "inappropriate" decisions. The rules also expand the list of types of cases to be excluded from public view: those involving state secrets, crimes committed by minors, divorce cases, cases involving custody or guardianship of children, and most disputes resolved through mediation. In addition, decisions may not be posted online for release until after the appeal process is exhausted, an area that was ambiguous under the 2013 rules.[15] Taken together these carve-outs undoubtedly restrict our view of significant areas of everyday adjudication, such as family law, as well as of how courts resolve difficult or sensitive disputes.[16] On balance, however, the 2016 interpretation significantly increases the range of court documents required to be made public.

In Henan, in particular, these two rounds of SPC rule making had crosscutting effects on the amount of publicly available information. In some areas, clearer curbs on judicial discretion expanded the public record. In the initial years after Henan courts began posting cases online, for example, courts were willing to hold decisions back when the parties objected. Today, it appears that they do so only when the case fits into a specific category set forth in the SPC rules. In other important areas, however, national standardization led to less disclosure. Henan courts routinely posted first instance decisions online, for example, until the 2016 rules explicitly placed them off limits until a case becomes final. In addition, divorce cases–which the Henan courts formerly posted unless one of the parties objected—are (at least in theory) uniformly offline under the 2016 SPC regulations.[17]

Why has China's judicial leadership embraced the practice of making court judgments public? Although the reasons behind China's embrace of mass publication of court decisions are complex, the primary motivation appears to be a desire to curb wrongdoing in the courts, rather than an interest in empowering individuals (or facilitating research by scholars). This line of argument was made explicit by court officials in Henan early in the process of putting cases online, and also appears to be influencing the SPC. There is a CCP tradition of harnessing the energy of crowds to reach government goals. Judges are more likely to follow the law, in other words, and less likely to engage in malfeasance, when they know their work will be made public. A 2017 SPC white paper endorsed this logic, noting that placing

14. No public posting of case number or explanation of the reason for non-posting is required for cases involving state secrets or national security. Nevertheless, because case numbers run sequentially by year in individual court decisions, it should be far easier in the future to identify the number of cases being held back without explanation.

15. The rules do, however, require that first instance decisions be made public alongside the appeal when the appellate decision is made public. Prior to the issuance of the 2016 SPC rules, there was some debate over whether non-final first instance court decisions should be made public. Those opposed to the idea were concerned that litigants might be confused or angry if decisions published online were later altered or reversed, and also that first-instance judges might face undue pressure from litigants.

16. One recent study of the SPC database found that high profile cases are often omitted (Ma, Yu, and He 2016, 222).

17. In informal discussions, however, judges say that actual practice varies from court to court.

cases online fits with President Xi Jinping's calls for judicial openness and increased public supervision (SPC 2017).

An additional goal of court efforts to make public tens of millions of court decisions is to raise the status of courts within the party-state, and with the public. Chinese courts have long been regarded as weak actors, and greater transparency may help improve trust in courts and make it easier for them to resist external pressure.[18] Court leaders closely control this strategy, however, by focusing attention on the outcomes of court cases rather than transparency of the legal process. Even though select videos of court proceedings are an increasingly common feature of court websites, media and public access to court proceedings remains limited.

Technophiles inside the political-legal system have also argued that algorithms derived from mass digitization of court opinions may facilitate greater efficiency and standardization among China's courts. Some court leaders are on the record discussing hopes that artificial intelligence can ensure consistent decision-making (同案同判), and even reduce judges' workload by drafting parts of opinions or deciding easy cases. SPC President Zhou Qiang, in particular, is associated with the idea of smart courts (智慧法院) and has talked about how computer-assisted judging could improve litigant satisfaction by ensuring consistent, fair, and transparent dispute resolution.[19] In contrast to past concerns about catching up with other countries, there is a possibility that Chinese courts could leapfrog past the rest of the world into the futuristic world of computerized judging. Although China's first forays with computer-assisted judging have been small-scale, and some Chinese judges express skepticism about computerized adjudication, these early experiments place new importance on court opinions as the raw source of data programmers would use to build software capable of computerized decision-making.

For now, though, the evolving world of online legal information is characterized by both the rapid expansion of available information and the instability of the platforms that host it. The SPC website, the case database most likely to attract researchers because of the government imprimatur, is particularly unstable. Different ways of searching the website yield different results,[20] and a fifteen-week effort by one of our research assistants to track the website's count of documents available for each province revealed that documents oc-

18. Other Chinese institutions, notably the Ministry of Environmental Protection and the stock exchanges, have also used the threat of public exposure to curb wrongdoing and boost their own standing. The difference is that environmental and securities authorities have sought to use transparency to control behavior of third parties, while courts are using transparency to control misconduct within their own institution. On how transparency is "good medicine" (良药) to combat favoritism and local government influence over judicial decision-making, see (Wei 2013).

19. See (Jie 2016)

20. The number of documents available depends on how the researcher sets her search parameters. To narrow down the search to include only documents from 2015, the researcher can set the dates of interest through the 高级检索 feature at the top of the webpage (Method 1). Alternatively, the researcher can use the filters on the left side of the webpage by clicking on "2015" under 按裁判年份筛选 (Method 2). Over the course of fifteen weeks, Method 2 always yielded fewer documents than Method 1. This discrepancy across search methods can likely be traced to differences in the back-end code.

casionally disappear.[21] In addition, the website has been overhauled twice, and the 2016 reboot appears designed to repel web crawlers and automated downloads.[22] All of these factors make data permanence a crucial concern, and underscore the urgency for scholars to preserve offline the documents that underlie any analysis.

3 Data and Methods: The Henan Dataset

The Henan dataset is a collection of 1,058,990 court documents downloaded from the Henan High Court website. We chose to focus on cases from a single province, Henan, for three main reasons. First, Henan started putting cases online earlier than most other provinces, and posted hundreds of thousands of cases prior to the launch of the SPC website. Studying Henan allows us to look further back in time than would be possible in other provinces, and, in future research, potentially to compare documents collected from the provincial high court website with what was subsequently posted to the SPC website. Second, Henan ranks in the bottom third of Chinese provinces in per-capita GDP. Examining court practice in Henan is a useful corrective to scholars' tendency to focus on courts in rich areas, where researchers have often enjoyed better access. Our study also reveals significant differences even within Henan–which is perhaps not surprising given that the province is home to nearly 100 million people and 184 courts. Third, a provincial focus makes finer-grained analysis possible. It is feasible to collect information about individual courts, as we do in the next section, and to explore reasons for differences among them.

Creating this freestanding collection of court documents took 18 months of effort by a team of research assistants with a background in computer science. Neither creating nor analyzing this database was easy and, whenever possible, we recount roadblocks in the hope our experience can ease the way for others. The team developed software to scrape hundreds of thousands of court documents,[23] and to parse those documents into components for analysis. This second step involved writing code known as a parsing script to help the computer differentiate different parts of court documents. Due to the number of documents involved, a parsing script is necessary for such basic tasks as counting document types or knowing how many documents each court uploaded.

21. One of our research assistants tracked the number of documents available on the SPC website over a fifteen week stretch in Fall 2016. Although the number of cases available for each province generally increased weekly, there were two exceptions: 1) The number of 2015 documents available for Heilongjiang decreased by 25 between November 18 and November 27, 2016; and 2) the number of documents available for the Xinjiang Production and Construction Corps decreased by 1 between November 12 and November 18, 2016. This could be due to a technical bug, or the purposeful removal of documents from the SPC website.

22. One possibility is that this reflects a desire to monetize the market for legal information. The SPC has begun selling access to its data to a select group of companies for a modest fee, and has also created an affiliated company under the SPC's Information Center in an apparent attempt to compete directly with the consulting services offered by some of the existing commercial websites. Future work will explore how public and private actors are jointly shaping the evolving market for legal information.

23. Web scraping requires writing a set of detailed instructions so that a computer can find and download cases as a human user might, but at a faster pace than humans can click. A balance needs to be struck so that the download rate is fast enough to preserve data without the activity overwhelming the website.

Writing a functional parsing script was our biggest technical challenge, largely due to variation in local court formatting of court decisions. The case identifier, in particular, is a critical piece of information that requires deep knowledge of court procedure to interpret and was also not yet standardized when we began. The case identification number (case ID) appears in the header of each document, and is a shorthand combination of numbers and characters that indicates the name of the court, the year the case was filed, the type of case, the case number, and the procedural posture of the case (usually first instance, appeal, rehearing, or enforcement).[24] SPC rules standardizing court practice for case identifiers and numbering came into effect in 2016. Previously, court practice varied widely within provinces and sometimes within individual courts.[25] For example, some Henan courts used one character to identify all civil cases while others used different characters to signify traffic, environmental, or financial civil cases. Our parsing script identified 669 variations in the type of case portion of the case ID.

In terms of format, court opinions follow guidelines set by the SPC, with some local variation.[26] They open with a header that includes the court name, case type, and case number. Next comes a list of parties to the case, including information about the lawyers, legal workers or other persons acting as legal representatives involved in the proceedings. The substantial middle of the opinion follows, with a summary of the claims and arguments presented by each party, the facts and evidence reviewed by the court, and the court's legal reasoning and decision. The final paragraph of the decision apportions legal fees to the parties, before closing with the names of the court personnel who heard the case and the date of the judgment.

The ability to parse cases successfully yielded Table 1,[27] a detailed look at the types of documents contained in the Henan dataset. The 1,058,990 court documents span the years between 1996 and 2015, with the great majority dating from 2008-2015. Major categories include 693,751 decisions in civil disputes, 255,255 decisions in criminal cases, and 31,710 decisions in administrative cases. There are also a substantial number of enforcement actions (54,192).

24. In theory, cases are numbered sequentially within court divisions each year. This means there should be a case numbered 1 in the civil, administrative, and criminal division of each court each year, followed by a case numbered 2 and so on, through the end of the year. Some courts have multiple civil or criminal divisions, and will generally number cases sequentially within each of these divisions.

25. For example, the case ID "(2014) 西民初字第926号" indicates a 2014 first instance (初) civil case (民) from the Xiping basic level court (西), with case number 926. This is the standard format for a case ID prior to the 2015 regulations (SPC 2015b).The SPC rules mandating standardized formatting will greatly aid future parsing algorithms because they assign a distinct character-numerical identifier to each court (eliminating the possibility of overlap of the court identifier portion of a case id) and standardize the characters used to indicate type of case, although anticipating the flow of natural language through a finite set of concrete rules will always be difficult.

26. See (SPC 2016b) for the most recent guidelines.

27. A few notes on classification: 1) Joint civil/criminal cases (刑事附带民事) are classified as criminal; 2) Administrative litigation enforcement actions are counted in enforcement actions, rather than administrative litigation; 3) The""unspecified" row contains the 16,378 documents where the date was missing or obviously wrong (e.g. year 2022).

Table 1: Documents in Henan Dataset

Year	Civil Decisions	Criminal Decisions	Administrative Decisions	Enforcement Decisions	Other	Total
2015	148025	39609	7913	26134	1228	222909
2014	194616	59609	9067	20926	1883	286101
2013	120050	46226	3884	4240	666	175066
2012	66317	30578	2429	972	201	100497
2011	59379	29751	2513	417	222	92282
2010	60714	31331	3187	428	349	96009
2009	42583	17779	2566	407	3363	66698
2008	1836	347	142	21	434	2780
2007	86	6	0	12	28	132
2006	19	2	0	0	13	34
1996-2005	65	16	5	10	8	104
Unspecified	61	1	4	625	15687	16378
Total	693751	255255	31710	54192	24082	1058990

Beyond these major categories, the dataset also includes many types of documents that are not decisions in administrative, civil, criminal, or enforcement actions, grouped in the table into the "other" category. The sheer range of documents available is surprising, and illustrates the wide range of roles courts play. For example, the Henan dataset includes 115 letters from courts to other government agencies, known as han (函), which are a form of official communication designed to help different parts of government coordinate their work. Many of these letters refer cases of criminal conduct uncovered through civil litigation to the police. The dataset also includes court orders subjecting individuals to compulsory medical treatment (强制医疗决定书)—usually decisions to commit a criminal defendant deemed mentally ill—and 1,497 rejections of post-judgment petitions (申诉通知书). Still other documents show courts reducing sentences for incarcerated defendants, using their administrative powers to detain individuals, and deciding state compensation claims. Judges everywhere inhabit a paper-dense world, and one noteworthy aspect of China's new transparency regime is how much of this paperwork it renders legible to others. This is the first time these documents have been made accessible to scholars, at least outside of China, and these everyday records offer a rich source for future work on the roles Chinese courts play.

4 What's Missing? Assessing Bias

For all that is available, the Henan dataset is also clearly incomplete. Charting the gaps in what Chinese courts choose to disclose is an urgent task for researchers, especially because existing work based on court documents culled from online databases rarely dwells on what is missing, and sometimes even fails to acknowledge the possibility of bias. Below, we document what we call "the missingness problem:" variation in court compliance with the national mandate for disclosure. We find wide variation in disclosure rates across courts that cannot be fully explained by differences in mediation rates, court level, GDP per capita, or

population.

Of course, figuring out what is not available is far from easy. As a starting point, Henan courts reported completing 685,890 cases in 2014, according to internal court statistics, compared to 285,382 documents in our collection. On average, this means that Henan courts placed just over 41 percent of their docket online, a proportion in line with recent national estimates that slightly less than half of 2014 and 2015 cases appear on the SPC website (Ma, Yu, and He 2016). However, this average disguises tremendous variation. The highest-ranking court in our sample released enough documents to plausibly cover 83 percent of completed cases, compared to just 14 percent for the least compliant court.[28] This variation is particularly striking for a province with detailed rules concerning what must go online that also ranks courts based on their compliance with provincial transparency policy.

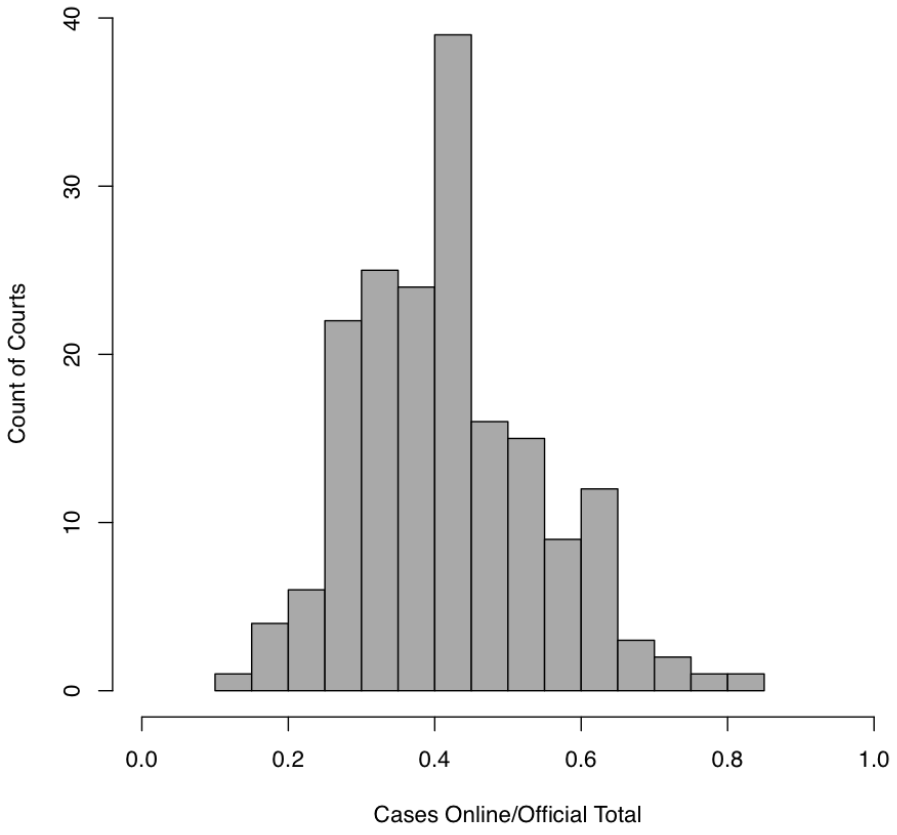

Figure 1: Proportion of Cases Online by Court in the Henan Dataset, 2014

At first glance, the most likely explanation for "the missingness problem" is variation in

28. All of the estimates should be treated as upper bounds. As discussed further below, courts sometimes publicly post more than one decision related to a case.

mediation rates. Many civil cases are resolved through mediation rather than adjudication, and mediation decisions are not posted online. In Henan, internal court statistics count 110,134 mediated cases in 2014.[29] Individual court mediation rates in civil cases in 2014 ranged from 11 to 78 percent of all cases with an average rate of 31 percent.[30] Figure 2 shows how accounting for the mediation of civil disputes, as described in footnote 32, shifts the proportion of cases online. Overall, the average proportion of court cases online rises to 50 percent after accounting for mediated cases, but there is even more variation across courts. Although the least compliant court only released enough documents to plausibly cover 20 percent of completed cases, there are now four courts that released enough documents potentially to cover 100 percent of decided cases. Adjusted for mediated civil cases, the most compliant court, the Sheqi county basic level court, released 9 percent more documents than the number of cases it reported completing in 2014.[31]

Court level strongly affects disclosure rates (see Figure 3). The average intermediate court released enough documents to potentially cover 70 percent of cases decided, compared to an average of 50 percent for basic level courts, a statistically significant difference.[32] What explains this pattern? One reason why intermediate courts and basic level courts diverge in their disclosure rates is likely the combination of docket composition and SPC rules governing which cases should be made public. Under SPC rules, first instance decisions may only be released after the case becomes final, meaning either that no appeal is filed within a stipulated period or a higher court decides the appeal. This creates a time lag between first instance decisions and putting those cases online. If judges are unaware when appeals are resolved, some first instance cases may not be posted online.[33] In contrast, the vast majority of intermediate court decisions are routine appeals, which courts may post immediately after the decision is issued.

The presence of outliers is also worth noting. Zhoukou city intermediate court is one the

29. We obtained data on the total number of mediated cases for nearly all Henan courts. This analysis is based on data from 180 courts, rather than all 183 courts in existence in 2014, as we were able to verify that three courts had clerical errors in the official data for mediation. Note that Henan established one new court in 2016, for a total of 184 courts as of early 2017. In this paper, we look at variation in missingness on the individual court level. Another approach is subtract the number of mediated cases (110,134) and the number of documents in the Henan dataset (285,382) from the total number of cases those 180 courts reported completing in 2014 (684,471). This yields a gap of 288,955 missing documents.

30. To account for mediation, we subtracted the number of mediated cases from the total number of cases the court reported handing. This gave us the total number of cases resolved through adjudication, which we compared to the number of documents for each court in the Henan dataset.

31. Courts may release multiple decisions in the same case, in order to resolve procedural or jurisdictional issues before issuing a final judgment. In the entire Henan dataset, there are only 199 duplicate case IDs, suggesting that there are few duplicate documents.

32. Using a two sample t-test, p=.0033.

33. This was a particularly important issue in 2014, when the rules about whether first instance cases should be placed online were unclear. It is likely a less of an issue today. As of 2016, first instance cases are supposed to be made automatically public after appeals are decided. In 2014, courts also courts also had widely different approaches posting the short judgments known as caidingshu (裁定书) that resolve cases without deciding the merits (for example, confirming the withdrawal of a civil claim by a plaintiff). Likewise courts varied in whether they put a range of documents relating to enforcement online.

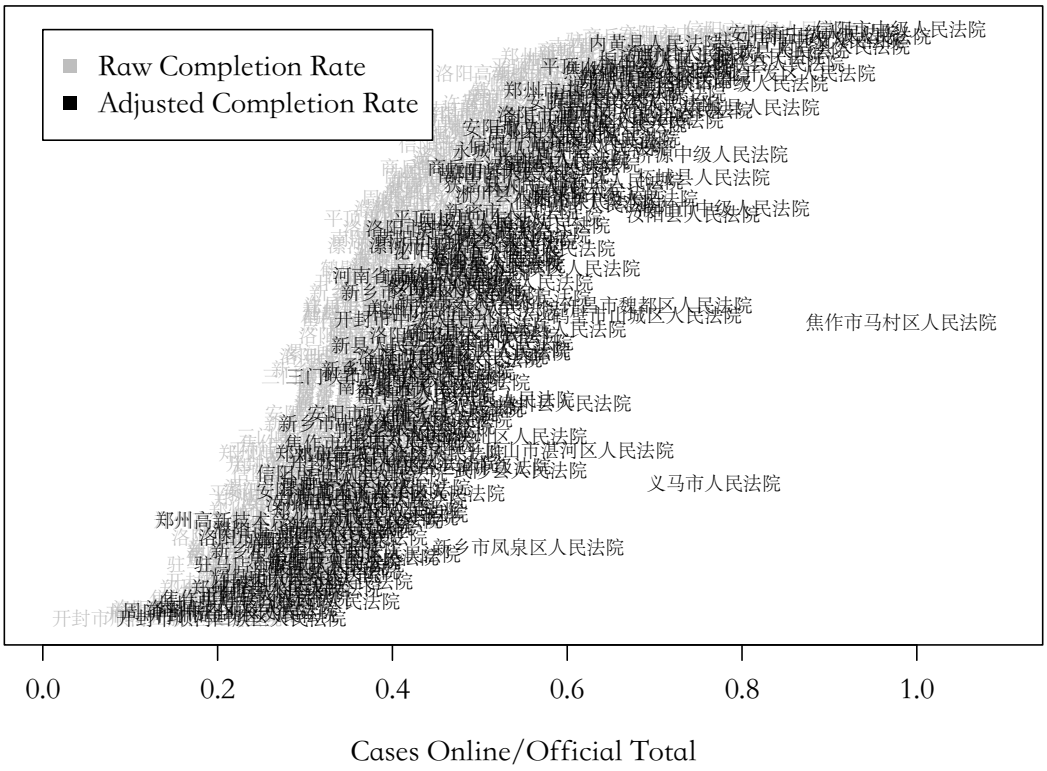

Figure 2: Proportion of Cases Online in the Henan Dataset, Adjusted for Mediation Rate

least transparent courts in our sample, while the Sheqi county basic level court is the most transparent. These findings set the stage for future research. What distinguishes these two courts from their peers? More work, both qualitative and quantitative, is needed to better understand variation in disclosure rates between courts, as well as between different levels of the judicial hierarchy.[34]

In particular, future work should explore how resource bias and incentive bias skew the available pool of court documents (Grimmer, Roberts, and Stewart, n.d.). Resource bias suggests that variation in court transparency stems from underlying resource constraints, particularly the availability of personnel to collect judicial decisions, blackout personal information, and place them online. For example, conversations with judges in Henan suggest that busier courts, especially those undergoing rapid growth in caseloads, may have been unable to prioritize placing cases online. As a first step toward investigating the existence of resource bias, we matched courts with GDP per capita and population (see Appendix B).

34. A further issue is that courts vary in how they collect and count cases, and some of the variation we observe might reflect differences in counting practices, or an inaccurate final tally in court statistics.

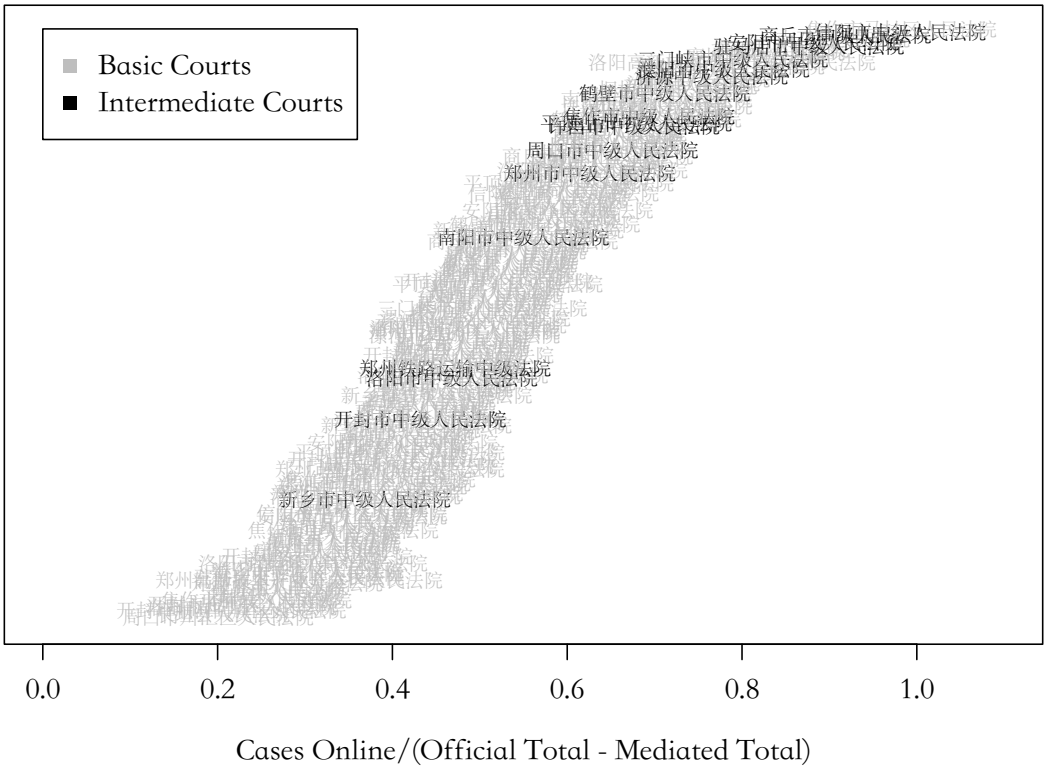

Figure 3: Basic Level Courts Versus Intermediate Courts, Proportion of Cases Online

There is no statistically significant relationship between these two variables and the proportion of the docket placed online, indicating either that resources are not strongly correlated with judicial disclosure rate or that GDP per capita and population are not a good measure of resources. Further exploration of research bias will also benefit from interviews with court personnel about how the process of document disclosure works, and whether those responsible for the task have enough time to devote to the work.

In contrast, incentive bias suggests that decisions about court transparency hinge on the incentives of the court leadership. There is already evidence, for example, that uploads of judicial decisions to the SPC website spike at the end of the quarter just ahead of court evaluations (Ma, Yu, and He 2016, 204).[35] Court presidents likely also assess incentives differently, and vary in the value they place on transparency as an idea as well as how much they care about their court's performance in annual province-wide transparency rankings. Courts appear to have adopted different practices regarding the degree to which they made compliance with requirements that cases be posted online part of the evaluation of individual judges. Some judges may even strategically place additional documents online to help their court's standing in the disclosure rankings. For example, some courts in Henan continued to release divorce decisions even after the SPC deemed divorce-related documents unsuitable for online publication. Likewise, in 2014 some courts posted first instance decisions as soon as they are issued despite SPC guidance to wait until the appeal process is exhausted. In a system that relies heavily on quantitative metrics for evaluating courts, over-releasing some types of documents may help court leaders avoid criticism for failing to comply with transparency policy.[36]

Variation in disclosure rates between courts is likely to become less extreme over time, as rules becomes settled about the types of documents subject to compulsory disclosure. In addition, the process for disclosing court documents has also become both automatic and uniform. In mid-2016, all Henan courts began using software capable of instantly sending completed decisions to the back-end of the SPC website. The default is also now to automatically clear cases for disclosure, unless the presiding judge indicates a specific reason the case should not be made public. All this suggests that the misssingness problem is poised to become less pronounced over time, although Chinese judges and academics comment that variation in court disclosure rates is unlikely to disappear entirely.

What is needed is a sustained, cooperative effort to document the types of biases in China's new online case databases and trace how those biases have changed over time. In fact, Judges in Henan express support for both the resource explanation and the incentive theory, saying that some court presidents emphasized document disclosure more than others, and also that some variation likely reflects the presence (or absence) of dedicated, competent staff whose job it was to place cases online. But what to do in the meantime, as missing cases will affect every analysis? The first wave of scholarship has focused on quantifying totals:

35. (Ma, Yu, and He 2016) discovered through fieldwork that some courts face a "small exam" each quarter (季度小考), as well as a major year-end evaluation (年度大考).

36. Of course, this behavior also could be due to administrative convenience or clerical errors.

the number and types of cases online, or similar counts of particular types of cases. For this type of approach, the main lesson is humility. Numbers need to be treated as estimates, and conclusions as conditional in recognition of the fact that the public record is partial. Researchers may also want to seek out pockets of excellent data, such as the seven courts in our dataset that place upwards of 90 percent of documents online. At least for now, a smaller scale analysis of a more complete corpus is a worthwhile tradeoff.

In the next section, we also suggest refocusing analyses of the court cases from quantifying totals to describing the text itself. In particular, records that document the sheer variety of legal disputes can surface unrecognized tensions in Chinese society, and offer insight onto how Chinese citizens interact with each other and with the state. Close reading will be central to this approach, though computational social science can certainly aid discovery by sifting and sorting documents. Topic modeling, a technique of unsupervised machine learning discussed below, is a particularly useful tool for translating the vast quantity of text courts produce into an account of what they do.

5 Court Decisions as Data: Insights into Administrative Litigation

To illustrate how topic modeling can reveal patterns in court activity and identify new lines of inquiry, this section focuses on 20,321 decisions by Henan administrative court divisions in our dataset.[37] To the best of our knowledge, no prior scholarship has analyzed such a large collection of Chinese administrative cases, and only a few scholars have used topic modeling to analyze judicial decisions in other countries (Livermore, Riddell, and Rockmore 2016).[38] Administrative law is also a natural starting point for students of Chinese politics because it is intrinsically political. This is the area of law that governs interactions between citizens and the state, and administrative lawsuits are often framed as a way for individuals or groups to protest government action.

Both inside and outside of China, observers closely track the number of administrative lawsuits as a metric for government accountability and official tolerance for legal challenges. Yet focusing on statistics – in particular, official national statistics – yields only modest benefits. National statistics break down administrative lawsuits into twelve general categories that give little insight onto the underlying claim or types of parties involved.[39] In contrast,

37. We begin with 20,321 of the administrative decisions because these were first posted online. As reflected in Table 1, we have subsequently collected an additional 11,389 administrative decisions and are in the process of updating our analysis based on new data.

38. Our approach also contrasts to recent empirical work on administrative litigation (Zhang, Ortolano, and Lü 2010; Zhang and Ortolano 2010; Cui, forthcoming; Li 2014), which draws on interviews and small-n samples of court documents. To be sure, the scale of administrative litigation is dwarfed by civil and criminal cases, both in our dataset and in the overall court system, and future work will examine other areas of law. Of the 13.9 million cases Chinese courts decided in 2014, just 130,964 were administrative cases (China Law Yearbook 2015; China Law Yearbook 2014).

39. These twelve categories are: urban construction, resources, public security, labor and social benefits, township government, traffic, business, family planning, public health, agriculture, tax, and other. Most land-

the basic and intermediate court decisions that make up the Henan dataset offer a far more detailed view of who uses administrative law, what topics end up in court, and how administrative law judges spend their time.

Topic modeling, a tool that originated in computer science and is now commonly used across the social sciences, helped us make sense of over twenty thousand documents without reading each one. Rather than replicate existing (and extensive) explanations of topic modeling,[40] we offer only a brief, layman's introduction here. We used the Structural Topic Model package in R (Roberts et al. 2014; Roberts, Stewart, and Tingley 2016; Roberts, Stewart, and Airoldi 2016), to estimate topics, which are groups of words that are likely to appear together within documents. For example, the model estimates that the words "land," "government," "villager," "dispute," and "to handle" frequently appear together, and form a topic that we labeled as "rural land rights." In addition, the model estimates topic proportions for each document, or the amount that each topic appears within the document.[41] Each document consists of a mixture of topics. For example, our topic model estimates that a 2014 decision by the Shaoxian county court on forestry rights can be described as partly topic 5 (rural buildings disputes), partly topic 12 (rural land rights), and partly topic 48 (jurisdiction). With assistance from Chinese administrative law scholars, we reviewed the highest frequency words associated with each topic, and the fifteen example documents that the model estimated to be best examples of the topic.[42] We then manually assigned a topic label to each topic (e.g. "birth planning," or "case withdrawals by natural person"). A list of all fifty topics, and the highest frequency words associated with them, appears in Appendix A.

One feature of the Henan dataset, which is also reflected in the topic model, is that it contains multiple types of administrative legal action. The two main categories are lawsuits against the state and non-litigation enforcement cases brought by government entities (非诉行政执行案子).[43] The conventional wisdom treats this first category of administrative

related disputes probably fall under urban construction, which comprised 17 percent of the administrative lawsuits filed in 2014 (China Law Yearbook 2015, 1015).

40. For a more extensive introduction to topic modeling, see (Blei, Ng, and Jordan 2003; Grimmer and Stewart 2013; Roberts et al. 2014). The highest probability words in each topic, as well as examples of cases in each topic, will be available in an online appendix.

41. A significant amount of pre-processing of the text is also required before running a topic model. In brief, we segmented the Chinese text into words using the Stanford Natural Language Processing Chinese segmenter, and removed stopwords and words that appeared in more than half of the documents. We chose fifty topics because we wanted to drill down several levels of detail beyond the twelve official categories of administrative litigation, and as a starting point to see how well the model helped us analyze the corpus of text. Other researchers might choose to generate fewer or more topics, depending on the desired degree of granularity. Since this model is used for exploration, we do not think there is one "correct" number of topics.

42. Close reading also helps validate the topics to ensure that they accurately reflect the content of the documents (Grimmer and Stewart 2013).

43. Article 15 of China's Administrative Compulsion Law gives administrative actors that lack the power of compulsory enforcement the ability to seek compulsory enforcement in court. In Henan, a third category of administrative decisions assign jurisdiction for hearing administrative cases. Other categories include administrative litigation enforcement actions, which are proceedings to enforce already decided administrative

litigation as part of "basic repertoire of contention" and a way for ordinary Chinese citizens to challenge government decisions (Mahboubi 2014; Cui, forthcoming). The Henan dataset does contain many examples of citizens suing the state, a type of lawsuit known in Chinese as *min gao guan* (民告官). For example, documents show Henan residents suing over administrative detention decisions following fights, to reverse the license granted for the construction of a new power plant, and to demand more severe punishment for a repeat violator of the food safety law. However, administrative court divisions also spend a great deal of time handling state-initiated requests for help with enforcement, a use of administrative law often overlooked by English language scholarship.[44] For example, our dataset contains a cluster of enforcement actions initiated by birth-planning officials seeking to collect social compensation fees for non-compliant births. Enforcement cases filed by state actors are common in China, with national figures for such enforcement actions slightly eclipsing the number of administrative lawsuits filed in 2014 (H. He 2016b, 624).[45]

Four broad categories of cases emerged from the topic model: 1) property disputes; 2) disputes over fines; 3) disputes over benefits, labor rights or compensation for workplace injuries, and 4) documents related to court procedure. We assigned each topic to one of these groups, based on the most important theme in the fifteen documents with the highest proportion of the topic.[46] Figure 4 shows a correlations plot of the fifty topics color-coded along these four themes, where the node size is proportional to the amount the topic is discussed within the corpus of text.[47] In other words, larger circles indicate more common topics. Edges, or the lines connecting nodes, indicate a correlation between topics greater than .01. Correlation means that the topics are more likely to appear within the same document.

Visually, the three largest topics jump out of Figure 4: birth planning (topic 15), case withdrawals by natural persons (topic 30) and fines for illegal occupation of land (topic 42). In addition, groups of topics that are likely to appear together within documents also come into focus. Some topics, such as birth planning (topic 15), stand alone as an insular group with little textual overlap with other documents. In contrast, a group of topics related to our benefits, labor, and worker compensation theme are clustered in the upper left of the plot.

lawsuits.

44. The exception is (Zhang, Ortolano, and Lü 2010), which discusses how environmental protection bureaus in Hubei province strategically use enforcement actions. Our dataset includes both administrative lawsuits and non-litigation enforcement actions, even though these are counted separately in national statistics. As a practical matter, the two types of documents can also be difficult to separate as court administrative divisions often use the same case numbering system and document titles for administrative litigation and non-litigation enforcement actions. In our dataset, some courts used specific case ids to distinguish non-litigation enforcement actions from lawsuits against the state, but many did not.

45. (H. He 2016b, 624) shows non-litigation administrative enforcement actions peaking at more than three times the number of cases brought by citizens against the state in 2000, before declining modestly over the next fourteen years. In 2014, enforcement actions were somewhat higher than the total number of administrative cases. Although Chinese-language scholarship has described the ebbs and flows of the total number of enforcement action, little has been written about the types of disputes likely to wind up as enforcement actions.

46. We classified two topics, 42 and 38, as both property and fines. They are classified as property in Figure 4 but will be dual-colored in a future draft.

47. Topic number in the plot corresponds to the topic number in the Appendix.

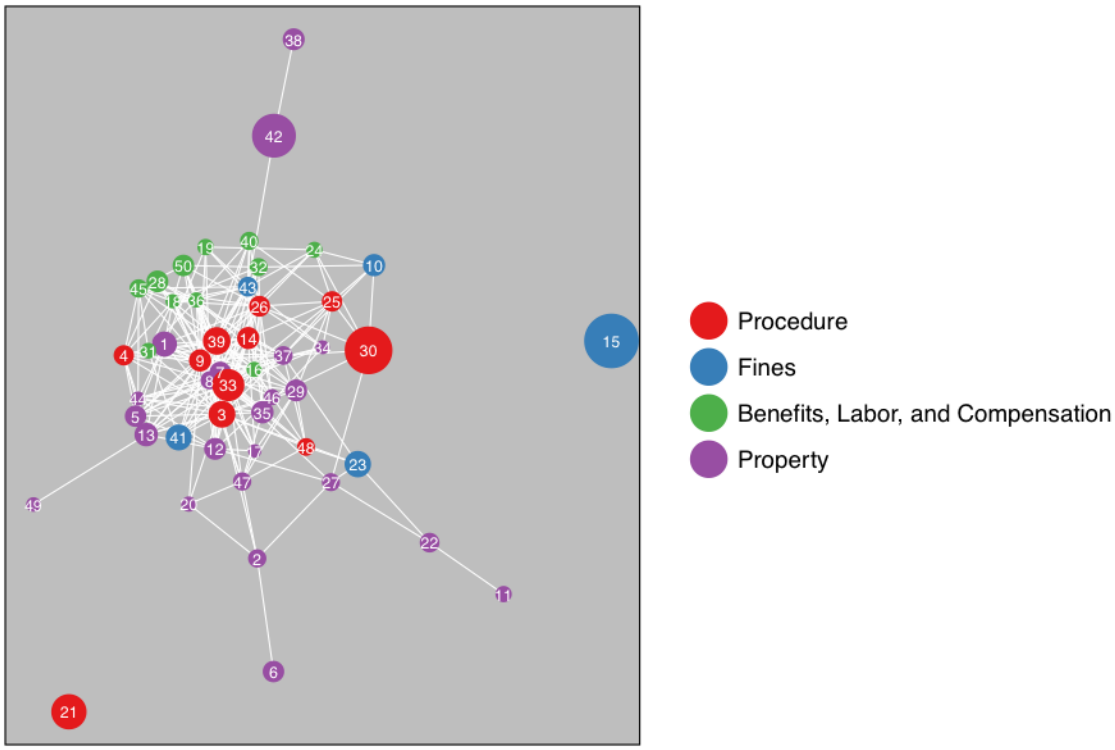

Figure 4: Correlations Plot

These documents are likely to contain similar language related to monetary demands, particularly claims related to health and retirement benefits. For example, topic 32 is primarily comprised of litigation over whether workers died on the job, such that their survivors are entitled to work-related death benefits. Property cases are clustered toward the bottom of the plot, with the exception of two land topics that are state-initiated enforcement actions to collect fines for illegal occupation of land or illegal buildings (topics 38 and 42). Language related to our themes of court procedure and fines appear in many types of cases, and are scattered throughout the correlation plot.

The most striking theme is the predominance of disputes concerning land and property, both rural and urban. Twenty-three topics involved land or property disputes, and 35 percent of all words in the corpus map onto one of the land topics.[48] As a large portion of the text in each court decision is related to court procedure and formalities, this means that a huge number of cases are related to property. For 64 percent of the documents in our dataset, at least one-third of the text is related to property.[49]

Although the missingness problem makes it impossible to estimate the percent of the

48. To estimate this, we summed the topic proportions of all property topics across all documents. This came out to 35 percent of the corpus.

49. In other words, the sum of property topics within each document is above 33 percent for 64 percent of the corpus.

total docket that deals with land, the topic model makes clear that property disputes are a hot topic. Even if none of the missing court decisions deal with property disputes—an extremely unlikely scenario—the property topic would have made up at least one-third of the Henan administrative docket in 2014.[50] On one level, this is not surprising given deep discontent over land takings in recent years. What is surprising is the range of property cases that arise: objections to the takings of property, disputes over compensation for taken property, registration of property to third parties, disputes over allegedly illegal rural buildings, and state efforts to enforce fines for illegal use of land. Even Topic 29, which initially looks like requests for information under China's open government information law, turns out to consist of cases brought by villagers facing land seizure. Similarly, Topic 37, which on first glance appears to be made up of complaints against the police, is actually dominated by cases brought by property owners suing the police over failure to protect their land from being seized by local officials or developers. Overall, the sweeping scope of property-related disputes suggest this issue dominates administrative law far beyond what national statistics report.[51]

Figure 5 shows the order of the property topics by their prevalence within the corpus, a ranking that helps demonstrate how far property disputes extend beyond government land takings. Topics about building takings and compensation, which we might have expected to dominate the corpus, rank in the middle of the list of topics, with cases about land takings and compensation near the bottom. In contrast, state attempts to enforce fines for illegal occupation of state land is the single largest topic, and comprises nearly 10 percent of words in the corpus. A closer look at these state enforcement cases, however, reveals a twist: Henan courts routinely reject government requests to help collect fines for the illegal occupation of buildings or land (topics 38 and 42). Although additional research would be needed to see if this pattern is typical, judicial reluctance to become involved in property disputes runs counter to the dominant narrative that Chinese courts side with the state to help it reach policy goals. Yet this behavior is consistent with the logic of responsibility avoidance that sometimes surfaces in research on Chinese courts, in which courts seek to sidestep controversy by finding for aggrieved citizens or by avoiding decisions in contentious cases (Liebman 2015, n.d.).

Topic modeling also provides a way to expand our mental map of administrative law beyond the two hemispheres of administrative litigation and state-led enforcement action. For example, there are five topics in which litigants seek to involve the state in what are primarily civil law disputes, a possibility overlooked in prior writing and one unobservable by looking at national statistics on the types of cases filed. The medical disputes topic, for example, consists largely of claims brought by patients against the health bureau for

50. The topic model estimates that 64 percent of the documents in the dataset are comprised at least one-third of property-related topics. If none of the missing court decisions contain any text related to property, an extremely unlikely scenario, then 33 percent of the true administrative docket would be comprised of decisions at least one-third related to property (.64*.52).

51. Likewise, our topic model almost certainly understates the preponderance of land-related disputes. In particular, decisions related to case withdrawals virtually never mention the underlying nature of the dispute, so any these documents that relate to property disputes will not be classified as property in the topic model.

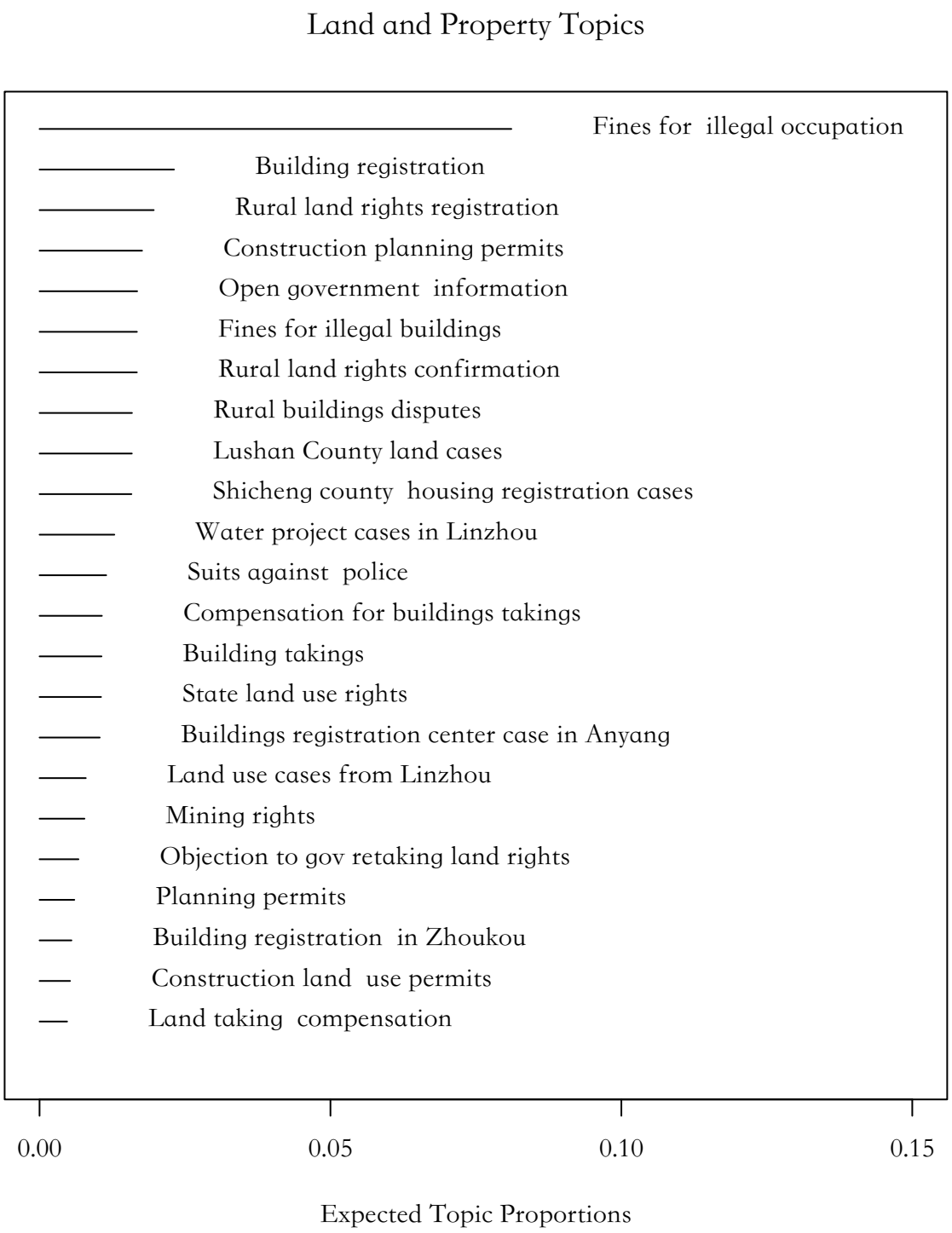

Figure 5: Prevalence of Property-Related Topics in the Corpus

failure to sanction doctors or hospitals (topic 36). In these cases, the underlying dispute is likely a civil malpractice suit against medical staff. Likewise, topic 16 is dominated by worker-initiated litigation against the local branch of the State Administration of Industry and Commerce objecting to the registration of the transfer of shares of their employer, the Luobei Heavy Machinery Company. The documents, however, suggest that the real issue was a labor dispute resulting from company restructuring. [52]The topic model reveals that litigants sometimes try strategically to transform civil disputes into administrative litigation, either to solicit state help resolving a problem or perhaps to demand state compensation. This trend is parallel to one observed in criminal law, where litigants at times try to use criminal sanctions to seek compensation in private disputes, reflecting continued difficulties enforcing civil judgments (Liebman 2015, 215-216).

A large group of topics also relates to court procedure, and this is clearly a large part of how Chinese administrative judges spend their time. Observers have long noted the high rate of withdrawals in administrative cases, and speculated that cases disappear from court dockets because of political pressure.[53] Our topic model confirms the prevalence of case withdrawals – 5 percent of words in the corpus are estimated to belong to one of the case withdrawal topics (topics 3, 26 and 30). Yet another 11 percent of words are estimated to belong to topics that cover routine issues, including jurisdictional issues (topics 23, 25, 48), appeals (topic 33), and evidence (39).[54] These other types of procedural documents suggest how much energy goes into the everyday work of running a court, as well as how much daily activity in an administrative law division does not relate to the resolution of substantive claims.

So what can topic modeling of Chinese cases teach us? First, it can help us question and refine accepted understandings. Many of the topics, particularly those related to property, lend support to the traditional notion in the English language literature that administrative litigation is a way for wronged individuals to protest government actions. Yet, even accounting for missing data, the model also highlights limitations to viewing administrative litigation solely through the lens of contentious politics. The state itself is an active user of administrative law, and a significant number of cases are efforts by citizens to involve the state in private disputes. Procedural decisions are also common, and there are many routine claims over topics such as retirement benefits or workplace injuries. Bringing the full spectrum of administrative law into view serves as reminder not to conflate activity inside administrative law divisions with increased demand for government accountability or to use administrative docket size as a measure of willingness of individuals to challenge the state or of courts' willingness to resolve such claims.

52. This is confirmed by the fact that multiple labor cases involving the same plaintiffs appear as civil disputes elsewhere in the Henan dataset.

53. (Mahboubi 2014) notes "a troublingly high percentage of cases are withdrawn... never less than 30 percent per year, reaching a high of almost 60 percent in 1997" (145). In 2015, the plaintiff withdrawal rate fell to 21.6 percent, the lowest level ever (H. He 2016a, 40).

54. These general topics do not account for the most general words in the corpus, as we removed words that appear in more than half of the cases before running the topic model.

Second, topic models excel at generating research questions. One striking—and puzzling—feature of our topic model is that twenty-two of the topics involve cases strings, defined as topics in which ten of the most representative fifteen examples relate to the same underlying dispute.[55] For example, the top fifteen examples in the traffic fines topic involve a single plaintiff protesting a series of fines levied by the Zhengzhou #10 traffic team (topic 10). The presence of so many related cases raises intriguing questions about the nature of administrative litigation. Do litigants who bring administrative litigation seek the protection of a crowd, organizing to spread political risk amongst a group? Or does the pervasiveness of case strings reflect judicial efforts to defeat collective litigation by breaking it up into individual claims, as (Stern 2013, 49) documents? Or perhaps are courts seeking to inflate their count of administrative cases by disaggregating claims, as they sometimes do to bump mediation rates (Li, Kocken, and Van Rooij 2016, 12)? Does the prevalence of so many case strings suggest that official statistics overstate the willingness of individuals to sue the state by counting what is essentially the same dispute numerous times? Do local government bureaus typically bunch enforcement actions together, perhaps to raise revenue or to signal a change in policy priorities?

Third, topic modeling also gives us a more textured sense of which government-affiliated actors use courts to help enforce their decisions (usually fines). In addition to the land-related enforcement actions discussed above, four other topics relate to state enforcement: lawsuits filed by the disabled persons association to collect fees on employers; by the local air defense bureau for overdue penalties related to contraction and land use; and by local governments for unpaid fines related to illegally constructed buildings or the illegal occupation of land. By definition, all of these actors lack legal authority to initiate compulsory enforcement of their decisions. However, this is a diverse group, whose membership includes both the well-known family planning bureau and more obscure players, such as the disabled persons association. Under what conditions, then, do state actors initiate court enforcement? As a longer chronological record builds up, it will also be possible to separate inveterate court users from the spikes in enforcement cases associated with the abrupt emergence of a new policy priority.

Fourth, notable silences in a topic model can also spark questions. In the Henan administrative topic model, for example, businesses bring very little litigation.[56] Does this suggest that companies have different ways of negotiating with local authorities, or that they are more content with the status quo? Lawsuits over pollution are also infrequent enough that they do not appear as a coherent topic.[57] Running a key word search on the word "pollution" (污染) pulls up just 98 documents in the administrative case dataset, of which only nine are lawsuits brought by citizens to protest air or water pollution. Does this reflect a province where environmental concerns are still secondary, or a calculated decision by Henan pollution victims that protest, petitioning, or civil litigation is more likely to yield results?

55. In some cases, the case string is a series of enforcement actions brought by the same government agency, in the same locality, over the same issue.

56. The exception is topic 26, which deals with case withdrawals by legal persons.

57. The vast majority of the pollution-related documents are NIMBY disputes, and consist in large part of topic 34 (construction land use permits). Administrative fines (topic 43), fines for illegal occupation (topic 42), and evidentiary disputes (topic 39) are other topics found in pollution-related documents.

Or are pollution cases systematically removed from the online record?

All of this makes topic modeling a useful tool for deepening our understanding of what courts do, even when working with incomplete data. The administrative topic model shows that this approach can be used for discovery and description of a dataset that is too large to read and code by hand. Topic modeling can uncover topics so mammoth that they remain large even accounting for missing cases, such as property disputes. It can throw out questions for future research, from case strings to notable silences. And it can surface underappreciated motifs, such as which state actors solicit judicial assistance with enforcement or citizens' creative efforts to draw the state into private disputes. These are not themes that would necessarily emerge from examining national statistics, tracking the news, or reading scholarship. Rather, they reflect the vantage point of the dataset itself: the perspective of the lower courts responsible for processing the diverse claims that collectively constitute administrative law. This wide-angle view offers a concrete sense of how this set of Chinese judges spend their time, and reminds us how much of administrative law falls outside the citizen versus the state frame.

6 Conclusion

For students of Chinese law and governance, the appearance of millions of court documents is simultaneously exciting and overwhelming. It is a leap to work at such a scale, and "the missingness problem" will affect every analysis until we arrive at a collective understanding of what is missing. At the same time, however, this new and rapidly expanding archive holds the potential to contribute to knowledge across a range of disciplines. Legal scholars, for example, may want to use parsed court documents to trace evolving interpretations of concepts such as fault, causation, or damages or to examine patters in how courts handle specific types of cases or parties. Political scientists could deploy topic models to visualize the rise and fall of certain topics in court dockets, and investigate the relationship to political priorities or legislative changes. There are also intriguing lines of interdisciplinary inquiry, such as investigating the institutional dynamics that led courts to embrace transparency or exploring whether gaps in the cases made public reflect limits in state capacity or administrative censorship. Indeed, a wave of digital scholarship on Chinese courts is already underway, as researchers both inside and outside of China flock toward a new source of information.

Methodologically, our work with the Henan database suggests several lessons. First, it is critical to take missing cases into account, rather than succumbing to the temptation to treat even a very large-n sample as an accurate reflection of reality. In particular, any calculation of the frequency of topics, arguments, or outcomes should be treated as an estimate, bounded by a discussion of how missing cases might matter. Second, viewing millions of court decisions provides an unparalleled wide-angle perspective on courts' daily activity, and exposes underlying patterns. Answering the deeper "why" and "how" questions, though, will likely continue to require the type of information about local context that typically emerges from close reading and time on the ground. Scholars must remember that court judgments provide only one, often limited, view of actual practice. Third, a migration toward treating

text as data in the field of Chinese law will require a multi-method approach that combines expertise and insights from law, the social sciences, and computer science.

Theoretically, the sudden availability of so much data from China's courts raises questions that our fields will be grappling with for years to come, especially because one of the insights of our initial effort is that the data itself can reveal fresh questions. Overall, however, is the availability of so much information about court judgments changing the practice of Chinese law? Does publication of court judgments encourage certain types of cases or legal arguments? Does the imperative to post court decisions affect court dockets, by making courts more or less willing to accept certain types of cases? Is greater transparency resulting in greater standardization and fairness in a legal system that has often been criticized as arbitrary and vulnerable to corruption? In this new environment, the list of possible research questions is nearly infinite and the primary challenge no longer is obtaining data but rather effectively using the public record, and prioritizing among the many questions thronging to be answered.

A Appendix A

	Topic Label	Category	Most frequent words, Chinese	Most frequent words, English
1	Building registration	Property	房屋,登记,权证,所有,房产	building, to register, warrant, all, real estate
2	Building takings	Property	征收,房屋,补偿,商丘市,政府	to levy (a fine), building, to compensate, Shangqiu City, government
3	Filing rejections	Procedure	河南省,副本,递交,中级,律师	Henan Province, duplicate, to hand in, middle level, lawyer
4	Statute of limitations	Procedure	起诉,期限,超过,知道,上诉人	to sue, deadline, to exceed, to become aware of, appellant
5	Rural buildings disputes	Property	宅基,政府,土地,集体,用地	homestead, government, land, collective, land use
6	Lushan County land cases	Property	新乡市,濮阳市,鲁山县,新密市,辉县市	Xinxiang City, Puyang City, Lushan County, Xinmi City, Huixian City
7	Shicheng county housing registration cases	Property	第三,证据,律师,具体,请求	third, evidence, lawyer, specific, to request
8	State land use rights	Property	土地,国土,政府,资源,使用权	land, state land, government, resources, usage rights
9	General topic	Procedure	没有,是否,具体,应当,存在	to not have, whether, specific, should, to exist
10	Zhengzhou traffic fines	Fines	交通,处罚,道路,违法,机动车	traffic, to punish, road, illegal, motor vehicle
11	Land use cases from Linzhou	Property	退耕,还林,粮食,补助,资金	restoring farmland to forest (tuigeng huanlin), food, subsidy, funds

12	Rural land rights confirmation	Property	土地,政府,村民,争议,处理	land, government, villager, dispute, to handle
13	Rural land rights registration	Property	土地,使用证,政府,颁发,登记	land, certificate of use, government, to award, to register
14	Challenges of reconsideration	Procedure	复议,政府,决定书,河南省,申请人	to reconsider, government, judicial decision, Henan Province, applicant
15	Family Planning	Fines	人口,生育,征收,计划,委员会	population, to give birth, to levy (a fine), plan, committee
16	SAIC registration objection	Benefits, Labor & Compensation	登记,变更,工商,企业,管理局	to register, to change, industry and commerce, company, administration bureau
17	Land taking compensation	Property	补偿,土地,征地,安置,征收	to compensate, land, to requisition land, to find a place for, to levy (a fine)
18	Health and worker benefits	Benefits, Labor & Compensation	信阳市,济源市,中心,煤矿,济源	Xinyang City, Jiyuan City, center, coal mine, Jiyuan (City)
19	Disability-related cases	Benefits, Labor & Compensation	灵宝市,抵押,镇平县,新郑市,盐业	Lingbao City, mortgage, Zhenping County, Xinzheng City, salt industry
20	Objection to gov retaking land rights	Property	改造,漯河市,政府,土地,征收	to transform, Luohe City, government, land, to levy (a fine)
21	Administrative enforcement actions (in Luyi)	Procedure	强制,鹿邑县,执行人,办公室,复议	to enforce, Luyi County, executor, office, to reconsider
22	Water project cases in Linzhou	Property	林州市,政府,土地,承包,批复	Linzhou City, government, land, contract, to reply

23	Police administrative penalties	Fines	处罚,公安局,分局,上访,管辖	to punish, public security bureau, sub-bureau, petition, jurisdiction
24	Taxi licenses	Benefits, Labor & Compensation	城市,汽车,出租,客运,车辆	city, car, to rent, passanger transport, vehicle
25	Procedural or technical issues	Procedure	郑州市,郑州,金水区,中原区,登封市	Zhengzhou City, Zhengzhou (City), Jinshui District, Zhongyuan District, Dengfeng City
26	Case withdrawals (legal persons)	Procedure	公司,有限,河南,集团,开发	company, limited, Henan (Province), group, development
27	Buildings registration center case in Anyang	Property	安阳市,长垣,正阳县,汝州市,林权证	Anyang City, Changyuan (County), Zhengyan County, Ruzhou City, forest warrants
28	Flood damage, Lvshi	Property	赔偿,损失,判决,卢氏县,请求	to compensate, loss/damage, judgment, Lushi County, to request
29	Open government information	Property	信息,公开,政府,答复,汤阴县	information, public, government, to answer, Tangyin County
30	Case withdrawals (natural persons)	Procedure	撤回,起诉,准许,律师,撤诉	to withdraw, to sue, to permit, lawyer, to withdraw
31	Retirement certificate disputes	Benefits, Labor & Compensation	某某,工作,人员,某甲,淅川县	so and so, to work, personnel, somebody, Xichuan County
32	Workplace death determination	Benefits, Labor & Compensation	工伤,认定,事故,工作,死亡	work injury, to determine, accident, to work, to die

33	Appeals	Procedure	上诉人,判决,上诉,原审,被上诉人	appellant, judgment, appeal, trial, appellee
34	Construction land use permits	Property	项目,建设,用地,规划,中牟县	project, to build, land use, to plan, Zhongmou County
35	Construction planning permits	Property	规划,建设,城乡,许可证,工程	to plan, to build, urban and rural areas, license, engineering
36	Medical disputes	Benefits, Labor & Compensation	平顶山市,鉴定,叶县,医院,医疗	Pingdingshan City, to identify, Ye County, hospital, medical treatment
37	Suits against police	Fines	公安局,分局,公安,机关,刑事	public security bureau, sub-bureau, public security, organ, criminal
38	Fines for illegal buildings	Property	土地,拆除,夏邑县,平舆县,非法	land, tear down, Xiayi County, Pingyu County, illegal
39	Evidentiary disputes	Procedure	证据,证明,提供,异议,提交	evidence, proof, to supply, objection, to submit
40	Counterfeit products rewards	Benefits, Labor & Compensation	工商,食品,处罚,管理局,举报	industry and commerce, food, to punish, administration bureau, report
41	Police public safety sanctions	Fines	处罚,公安局,治安,决定书,笔录	to punish, public security bureau, law and order, judicial decision, to record
42	Fines for illegal occupation	Property	处罚,国土,执行人,资源,决定书	to punish, state land, executor, resources, judicial decision
43	Administrative fines	Fines	处罚,违法,管理,执法,施工	to punish, illegal, to supervise, to enforce a law, construction

44	Building registration in Zhoukou	Procedure	上蔡县,周口市,面积,房屋,登记	Shangcai County, Zhoukou City, tract of land, building, to register
45	Elderly insurance	Benefits, Labor & Compensation	劳动,社会,单位,保险,保障	labor, society, work unit, insurance, to ensure
46	Mining rights	Fines	南阳市,南召县,方城县,焦作市,资源	Nanyang City, Nanzhao County, Fangcheng County, Jiaozuo City, resources
47	Compensation for buildings takings	Procedure	拆迁,房屋,开封市,补偿,政府	to demolish, building, Kaifeng City, to compensate, government
48	Jurisdiction	Procedure	洛阳市,授权,特别,代理,洛阳	Luoyang City, to authorize, especially, proxy, Luoyang (City)
49	Planning permits	Procedure	马店市,建设,土地,规划,许可证	Madian City, to build, land, to plan, license
50	Workplace injury	Benefits, Labor & Compensation	工伤,认定,劳动,工作,人社	work injury, to determine, labor, to work, person society

B Appendix B

In this Appendix, we show that the proportion of total cases online are not significantly related to either the GDP per capita of the locality or the population of the locality. This finding does not support the resource hypothesis that localities with more money or people would be better equipped to put cases online. Because intermediate and basic courts have significantly different rates of transparency, we divide the analysis into two sections – first to see if there GDP per capita and population is related to transparency in intermediate courts (Table 3, and then in basic courts (Table 4).

Table 3: Relationship between GDP per capita, population, and transparency in Intermediate Courts

	Dependent variable: Proportion of Cases Online		
	(1)	(2)	(3)
pop	−0.0003		−0.0001
	(0.0004)		(0.0005)
GDP.capita		−0.00000	−0.00000
		(0.00000)	(0.00000)
Constant	0.736***	0.908***	0.911***
	(0.071)	(0.143)	(0.148)
Observations	17	17	17
R^2	0.030	0.136	0.140
Adjusted R^2	−0.035	0.078	0.017
Residual Std. Error	0.176 (df = 15)	0.166 (df = 15)	0.171 (df = 14)
F Statistic	0.459 (df = 1; 15)	2.359 (df = 1; 15)	1.141 (df = 2; 14)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 4: Relationship between GDP per capita, population, and transparency in Basic Courts

	Dependent variable: Proportion of Cases Online		
	(1)	(2)	(3)
pop	−0.00003		−0.0001
	(0.0004)		(0.0005)
GDP.capita		−0.00000	−0.00000
		(0.00000)	(0.00000)
Constant	0.493***	0.502***	0.510***
	(0.029)	(0.026)	(0.043)
Observations	154	154	154
R^2	0.00004	0.001	0.002
Adjusted R^2	−0.007	−0.005	−0.011
Residual Std. Error	0.150 (df = 152)	0.150 (df = 152)	0.151 (df = 151)
F Statistic	0.006 (df = 1; 152)	0.213 (df = 1; 152)	0.131 (df = 2; 151)
Note:			*p<0.1; **p<0.05; ***p<0.01

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Cui, Wei. Forthcoming. "Does Judicial Independence Matter? A Study of the Determinants of Administrative Litigation in an Authoritarian Regime." *University of Pennsylvania Journal of International Law.* https://perma.cc/55QH-UPE7.

Evans, James A, and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory." *Annual Review of Sociology* 42:21–50.

Grimmer, Justin, Margaret E. Roberts, and Brandon Stewart. n.d. "Text as Data: Identifying, Acquiring, and Representing Text Quantitatively." *Unpublished Working Paper.*

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political analysis:* 267–297.

He, Haibo. 2016a. "How Much Progress Can a Legislation Bring? The 2014 Amendment of the Administrative Litigation Law of PRC." *Unpublished manuscript.*

———. 2016b. *Xingzheng susong fa* 行政诉讼法 *[Administrative Litigation Law].* Vol. 2. Beijing: China Law Press.

He, Weifang. 2003. "Jianshi touming fayuan 建设透明法院 [Creating Transparent Courts]." *Southern Weekend.* https://perma.cc/ZD5Z-HXL7.

He, Xin, and Yang Su. 2013. "Do the "haves" come out ahead in Shanghai courts?" *Journal of Empirical Legal Studies* 10 (1): 120–145.

Heilmann, Sebastian, and Elizabeth J Perry. 2011. *Mao's Invisible Hand: The Political Foundations of Adaptive Governance in China.* Harvard University Asia Center Cambridge, MA.

Hollyer, James R, B Peter Rosendorff, and James Raymond Vreeland. 2015. "Transparency, Protest, and Autocratic Instability." *American Political Science Review* 109 (04): 764–784.

Horsley, Jamie P. 2007. "China adopts first nationwide open government information regulations." *Freedominfo.org.* https://perma.cc/E5NR-QER3?type=image.

———. 2010. "Update on China's open government information regulations: Surprising public demand yielding some positive results." *FreedomInfo.org.* https://perma.cc/M9VE-FVLE.

Howson, Nicholas Calcina. 2010. "Corporate Law in the Shanghai People's Courts, 1992-2008: Judicial Autonomy in a Contemporary Authoritarian State." *East Asia Law Review* 5:303–442.

Huang, Philip C.C. 2017. "Dispatch Work in China:A Study from Case Records, Part II." *Modern China* 43:247–287.

Jie, Zhu. 2016. "Zhou Qiang: Jiakuai zuihui fayuan jianshe cujin sifa wei min gongzheng sifa 周强: 加快智慧法院建设促进司法为民公正司法 [Zhou Qiang: Speed Up the Creation of "Smart Courts" to Promote Justice and Fairness]." *Supreme People's Court.* https://perma.cc/6M9T-7QE4.

Li, Ji. 2013. "Suing the Leviathan—An Empirical Analysis of the Changing Rate of Administrative Litigation in China." *Journal of Empirical Legal Studies* 10 (4): 815–846.

———. 2014. "Dare You Sue the Tax Collector? An Empirical Study of Administrative Lawsuits Against Tax Agencies in China." 23 (1): 57–112.

Li, Yedan, Joris Kocken, and Benjamin Van Rooij. 2016. "Understanding China's Court Mediation Surge: Insights from a Local Court." *Law & Social Inquiry.*

Liebman, Benjamin L. 2005. "Watchdog or demagogue? The media in the Chinese legal system." *Columbia Law Review:* 1–157.

———. 2011. "The Media and the Courts: Towards Competitive Supervision?" *The China Quarterly:* 833–850.

———. 2015. "Leniency in Chinese Criminal Law? Everyday Justice in Henan." *Berkeley. J. Int'l L.* 33 (1): 153–222.

———. n.d. "Practical Justice? Ordinary Tort Litigation." *Unpublished Working Paper.*

Liebman, Benjamin L., and Tim Wu. 2007. "China's Network Justice." *Chi. J. Int'l L.* 8:257.

Livermore, Michael A, Allen Riddell, and Daniel Rockmore. 2016. "Agenda Formation and the US Supreme Court: A Topic Model Approach." *Virginia Law and Economics Research Paper No. 2.* https://perma.cc/YBB8-NMVZ.

Lorentzen, Peter, Pierre Landry, and John Yasuda. 2013. "Undermining Authoritarian Innovation: The Power of China's Industrial Giants." *The Journal of Politics* 76 (1): 182–194.

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-assisted Text Analysis for Comparative Politics." *Political Analysis* 23 (2): 254–277.

Ma, Chao, Xiaohong Yu, and Haibo He. 2016. "Da shuju fenxi: Zhongguo sifa caipan wenshu shangwang gongkai baogao 大数据分析: 中国司法裁判文书上网公开报告 [Empirical Analysis of China's Online Court Decision Database]." *Zhongguo Falv Pinglun* 中国法律评论*:* 195–246. https://perma.cc/6T5V-KNXH.

Mahboubi, Neysun. 2014. "Suing the Government in China." *Democratization in China, Korea, and Southeast Asia. Londres: Routledge:* 141–155.

Malesky, Edmund, Paul Schuler, and Anh Tran. 2012. "The Adverse Effects of Sunshine: A Field Experiment on Legislative Transparency in an Authoritarian Assembly." *American Political Science Review* 106 (4): 762–786.

Miller, Ian Matthew. 2013. "Rebellion, Crime and Violence in Qing China, 1722–1911: a topic modeling approach." *Poetics* 41 (6): 626–649.

National People's Congress Standing Committee. 2013. "Zhonghua Renmin Gongheguo Minshi Susongfa 中华人民共和国民事诉讼法endCJK* [Civil Procedure Law of the People's Republic of China']." https://perma.cc/MBX4-V2XX.

Roberts, Margaret E, Brandon M Stewart, and Edoardo M Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111 (515): 988–1003. https://perma.cc/MBX4-V2XX.

Roberts, Margaret E, Brandon M Stewart, and Dustin Tingley. 2016. "Navigating the Local Modes of Big Data: The Case of Topic Models." *Data Analytics in Social Science, Government, and Industry:* 51–97. https://perma.cc/MBX4-V2XX.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Chris Lucas, Jetson Leder-Luis, Bethany Albertson, Shana Gadarian, and David Rand. 2014. "Topic models for Open Ended Survey Responses with Applications to Experiments." *American Journal of Political Science:* 1064–1082. https://perma.cc/MBX4-V2XX.

Sina Court Channel. 2016. "Zhongguo Caipan Wenshu Wang: Quanqiu zuida wenshu gongkai pingtai 10 yue 1 ri qi shixing 大数据分析: 中国司法裁判文书上网公开报告 [China Online Court Judgments: World's Largest Online Judgments Platform to Come Into Effect October 1]." *Xinlang Fayuan Pindao* 新浪法院频道. https://perma.cc/9KQQ-HQB4.

Stern, Rachel E. 2013. *Environmental litigation in China: a study in political ambivalence.* Cambridge: Cambridge University Press. https://perma.cc/MBX4-V2XX.

———. 2014. "The Political Logic of China's New Environmental Courts." *The China Journal,* no. 72: 53–74. https://perma.cc/MBX4-V2XX.

Supreme People's Court. 2009a. "Renmin Fayuan Disange Wunian Gaige Gangyao 人民法院第三个五年改革纲要 [Third Five-Year Reform Outline for the People's Courts (2009-2013)]." https://perma.cc/MBX4-V2XX.

———. 2009b. "Zuigao Renmin Fayuan Yingfa Guanyu Sifa Gongkai de Liuxiang Guiding He Guanyu Renmin Fayuan Jieshou Xinwen Meiti Yulun Jiandu De Ruogan Guiding de Tongzhi 最高人民法院印发《关于司法公开的六项规定》和《关于人民法院接受新闻媒体舆论监督的若干规定》的通知 [Supreme People's Court's Notice on the Publication of Six Measures on Judicial Openness and Certain Provisions on People's Court Accepting News Media Supervision]." https://perma.cc/TT2U-J4WB.

———. 2013. "Zuigao Renmin Fayuan Guanyu Renmin Fayuan Zai Hulianwang Gongbu Caipan Wenshu De Guiding 最高人民法院关于人民法院在互联网公布裁判文书的规定 [Provisions of the Supreme People's Court on the Issuance of Judgments on the Internet by the People's Courts]." https://perma.cc/3A2L-QSVF.

Supreme People's Court. 2015a. "Zuigao renmin fayuan guanyu sheyong zhonghua renmin gongheguo minshi susongfa 最高人民法院关于适用《中华人民共和国民事诉讼法》的解释 [Interpretation of the Supreme People's Court of Several Issues concerning the Enforcement Procedures in the Application of the Civil Procedure Law of the People's Republic of China]." http://bit.ly/2sjn3t7.

———. 2015b. "Zuigao renmin fayuan guanyu yinfa 'guanyu renmin fayuan anjian de an hao de ruo gan gui ding' ji peitao biaozhun di tongzhi 最高人民法院关于印发《关于人民法院案件案号的若干规定》及配套标准的通知 [Notice of the Supreme People's Court on Printing and Distributing the Provisions on People's Court Cases and Supporting Standards]." https://perma.cc/7YV7-7SYB.

———. 2016a. "Zuigao renmin fayuan guanyu renmin fayuan zai hulianwang gongbu caipan wenshu de guiding 最高人民法院关于人民法院在互联网公布裁判文书的规定 [Supreme People's Court Regulations Regarding Placing Judicial Decisions on the Internet]." https://perma.cc/NG8N-BCJ6.

———. 2016b. "Zuigao renmin fayuan guanyu yinfa 'renmin fayuan minshi caipan wenshu zhizuo guifan' 'minshi susong wenshu yangshi' de tongzhi 最高人民法院关于印发《人民法院民事裁判文书制作规范》《民事诉讼文书样式》的通知 [Notice of the Supreme People's Court on the Issuance of 'Standards for Judicial Decision"' and 'The Style of Civil Litigation Documents']." https://perma.cc/LYC6-ZUWW.

———. 2017. "Judicial Transparency by People's Courts." https://perma.cc/MN9U-5K6L?type=image.

Wei, You. 2013. "Yong zhidu paichu quanli ganyu weihu sifa quanwei 用制度排除权力干预维护司法权威 [Use Institution to Prevent Political Intervention and Safeguard Judicial Authority]." *Legal Daily.* https://perma.cc/KBX2-QHDJ.

Zhang, Xuehua, and Leonard Ortolano. 2010. "Judicial Review of Environmental Administrative Decisions: Has it Changed the Behavior of Government Agencies?" *The China Journal,* no. 64: 97–119. https://perma.cc/MBX4-V2XX.

Zhang, Xuehua, Leonard Ortolano, and Zhongmei Lü. 2010. "Agency empowerment through the administrative litigation law: Court Enforcement of Pollution Levies in Hubei province." *The China Quarterly* 202:307–326. https://perma.cc/MBX4-V2XX.

National People's Congress Standing Committee. 2014. *Zhongguo falv nianjian* 中国法律年鉴 *[China Law Yearbook].* Beijing: China Statistics Press. https://perma.cc/MBX4-V2XX.

———. 2015. *Zhongguo falv nianjian* 中国法律年鉴 *[China Law Yearbook].* Beijing: China Statistics Press. https://perma.cc/MBX4-V2XX.