Mass Digitization of Chinese Court Decisions

HOW TO USE TEXT AS DATA IN THE FIELD OF CHINESE LAW

BENJAMIN L. LIEBMAN, Columbia Law School
MARGARET E. ROBERTS, University of California, San Diego
RACHEL E. STERN, University of California, Berkeley
ALICE Z. WANG, Columbia Law School

ABSTRACT

Since 2014, Chinese courts have placed tens of millions of court judgments online. We analyze the promise and pitfalls of using this new data source, highlighting takeaways for readers facing similar issues using other collections of legal texts. Drawing on 1,058,986 documents from Henan Province, we identify problems with missing data and call on scholars to treat variation in court disclosure rates as an urgent research question. We also outline strategies for learning from a corpus that is vast and incomplete. Using a topic model of administrative litigation in Henan, we complicate conventional wisdom that administrative lawsuits are an extension of contentious politics that give Chinese citizens an opportunity to challenge the state. Instead, we find a high prevalence of administrative cases that reflect an underlying dispute between two private parties, suggesting that administrative lawsuits are often an attempt to enlist help from the state in resolving an underlying civil dispute.

I. 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

We are grateful to a large number of commentators in both China and the United States whose feedback helped improve this article and to the many research assistants at Berkeley, Columbia, and the University of California, San Diego, who worked on various stages of this project. Particular thanks to Kevin Coakley, Subhasis Dasgupta, Amarnath Gupta, Haoshen Hong, and Kai Lin at the San Diego Supercomputer and Xiaohan Wu at Columbia Law for their help with the data. This work was partially funded by the National Science Foundation RIDIR program, award 1738411. Contact the corresponding author, Benjamin L. Liebman, at BL2075@columbia.edu.

Journal of Law and Courts (Fall 2020) 2164-6570/2020/0802-0001$10.00. Electronically published August 27, 2020.

led to a rapid expansion of the public record of court activity: more than 93 million documents were posted by June 2020.[1] 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 in treating legal texts as data and using computer science tools to uncover patterns in document collections too big for any research team to code by hand. Legal systems around the world are producing public text at unprecedented rates, and there is much to be learned from 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 legal scholarship. There is already a groundswell of interdisciplinary legal scholarship that applies computational text-analysis techniques to legal texts as diverse as Supreme Court and appellate court decisions (Bryan and Ringsmuth 2016; Livermore, Riddell, and Rockmore 2017; Rice 2017), amicus curiae briefs (Evans et al. 2007), transcripts of US Supreme Court oral arguments (Patton and Smith 2017), trial records from London's Old Bailey (Klingenstein, Hitchcock, and DeDeo 2014), 19th-century state procedure codes (Funk and Mullen 2018), and the world's constitutions, past and present (Law 2016, 2019).

This article stands at the intersection of these two trends, toward the increasing availability of legal texts on the one hand and computational social science on the other. Primarily, it is meant as a guide to a specific corpus: the millions of Chinese court decisions now online. This new source is impossible to ignore and can teach us a great deal, particularly as political conditions in China make interviews and surveys increasingly difficult. We provide information on how the corpus came about, use external data to reveal the holes in the data, and provide an example of how substantive conclusions can be drawn from the data set.

At the same time, however, few of the technical and methodological challenges we describe are unique to China. Any researcher looking to conduct computerized text analysis of court decisions would need to grapple with a similar list of issues: holes in the public record and little information about what is missing, changing rules on what courts should post and varying levels of compliance with the rules governing public availability, inconsistent formatting of court documents, the fluctuating nature of databases that continuously expand but also sometimes remove material, and the technological difficulty of large-scale downloads from websites not designed for that purpose. Our general call is for legal scholars to take documenting missing data in court decisions seriously and treat patterns of missingness as an important research question, even as we also sketch out approaches to learning from a corpus that is both giant and incomplete. This piece was written in anticipation of a coming wave of research that brings computerized text analysis to

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

the comparative study of courts, paralleling a similar trend in comparative politics.[2] Readers of the *Journal of Law and Courts* pursuing this methodology or tracking this turn in the field will be interested in our recommendations about how best to explore a vast, incomplete corpus.

To do this, we analyze a data set of 1,058,986 documents from the Henan Province High People's Court website.[3] Our analysis has two parts. First, we investigate how much of the court record is missing in 2014, finding wide intraprovincial variation in judicial disclosure rates that persists even after accounting for court level, mediation rates, gross domestic product (GDP) per capita, and population. Although we find that disclosure rates improved modestly by 2016, an average of 47% of Henan court decisions are nevertheless still missing. These findings underscore the risk of assuming that even a large case database is complete and also frame an important question for future research: how best to investigate competing explanations for missing data. Above all, these findings push back against the dominant approach in the deepening Chinese-language literature, which is to move quickly past problems with missing data and report frequency rates from an incomplete data set as if they represent reality.[4]

Although the scale and persistence of missing data complicate prospects for research, the second part of this article provides an example of how computational text analysis can yield insights even with data we know to be incomplete. A topic model of 25,921 decisions in administrative lawsuits from Henan shows how combining unsupervised machine learning with close reading of selected cases adds nuance to our understanding of how courts work. Since China introduced administrative litigation in 1989, most observers have treated it as a barometer of citizens' willingness to sue the state, even if a popular saying suggests it is as useful as "throwing an egg against a stone" (Finder 1989). Our findings push back against this conventional wisdom by showing that a significant portion of administrative litigation consists of attempts to draw the state into private disputes, a phenomenon poorly captured by either a contentious-politics frame or a government accountability frame. We use our recommended techniques for working with missing data to show that this "administrivization of private disputes" constitutes both a significant and underappreciated portion of administrative dockets, a finding that requires us to adjust how we think about Chinese administrative law.

II. THE ORIGINS OF JUDICIAL DISCLOSURE

Today's courts devote far more energy to judicial transparency than was the case a century ago. Worldwide, many courts selectively publish opinions, disclose court statistics, hold

2. For two overviews of how political scientists can use textual data, see Grimmer and Stewart (2013) and Lucas et al. (2015).

3. Our data set 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.

4. For a sample of articles that use this approach, see Cai and Liu (2016), Zhao and Geng (2016), and Wu and Zhang (2017).

press conferences, offer status updates on cases, maintain social media accounts, and allow television coverage of court proceedings. An emerging strand of the comparative literature on courts examines why courts go public and what effect transparency has on public trust and legitimacy. So far, however, this literature has focused on democracies, particularly in Latin America (Staton 2010; Ingram 2017) and Europe (Grendstad, Shaffer, and Waltenburg 2017). Certainly, government transparency is associated with democracies because sharing information seems intuitively linked to responsive, accountable government. Yet the trend toward judicial transparency is not limited to democracies, and one contribution of this article is to document how it extends even to authoritarian states such as China.[5]

When it comes to making court decisions available online, China is a trendsetter in the authoritarian world and is unusual even among other civil law jurisdictions. 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, until recently court decisions were typically available only to the people directly involved in the case. This meant that collecting even a small set of court opinions required personal contacts. Today, China's Supreme People's Court claims that the tens of millions of cases available online make its website the largest collection of public cases globally (Sina Court Channel 2016).[7] Vietnam, a Communist Party–led one-party state where political and legal reforms are often inspired by Chinese practice, has already followed suit. Starting on July 1, 2017, all Vietnamese courts became obligated to make their judgments public, with decisions centrally collected on a website managed by the Vietnamese Supreme People's Court (Nguyen 2019).

Why did China's judicial leadership embrace the practice of making court judgments public? One angle on this question is historical. As sudden as the 2013 decision to release the vast majority of court decisions might have seemed, it followed a history of experimentation and intellectual debate. The SPC recognized the value of making some court decisions publicly available to educate both judges and litigants early in the reform era. In 1985, the SPC began publishing an official gazette (最高人民法院公报), which included a small number of cases. Although not formally recognized as precedent, published cases were meant to guide lower courts on how to handle particular points of law and to improve the uniformity and quality of court decisions across China. Collections of representative cases (典型案例), some curated by the SPC and others by academics or local courts, also became standard fare for legal publishers and a source of practical guidance for their readers

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 2014; 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 review and read final court decisions, except for those involving state secrets or trade secrets or relating to personal privacy (National People's Congress Standing Committee 2012).

7. The claim is difficult to verify.

(Liebman and Wu 2007, 289). Likewise, in the late 1990s and early 2000s, individual courts began posting selected representative cases online. How much material was available varied, 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 LexisNexis and Westlaw, although coverage remained spotty. A few law firms also started posting cases handled by their lawyers to project professionalism and attract clients.

In the early 2000s, liberal scholars began to call for courts to place all opinions online as a way to fight corruption and restore public confidence in the courts (He 2003). Somewhat ironically, it was not until Wang Shengjun—an official widely perceived as ideologically conservative—became president of the SPC in 2008 that the SPC itself made a push to place large numbers of court decisions online. The SPC endorsed judicial transparency as a goal in its third 5-year plan for legal reform in 2009 and also encouraged lower courts' efforts to compile and publish judicial decisions (SPC 2009a, 2009b). These central government cues prompted local initiatives, including in Henan, where a mid-2009 Henan High People's Court order mandated that all courts in the province place the vast majority of decisions online. 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 People's Court President Zhang Liyong (Liebman 2015, 161–62). Other provinces followed Henan's lead. Building on these local experiments, the SPC under the leadership of SPC President Zhou Qiang called on courts nationwide to begin posting most cases online in 2013 (SPC 2013).[8] The new rules created a centralized website, called China Court Judgments (中国裁判文书网), which launched on July 1, 2013. By the middle of 2015, the website included documents from across the country (Sina Court Channel 2016).

What the SPC has embraced is a form of "controlled transparency" (Liebman 2011, 847). Although the trend is toward releasing a greater number and variety of court documents, certain types of cases are exempt, and local courts were granted discretion to hold back individual decisions. For example, the first set of SPC rules governing the public release of court opinions, issued in 2013, provided exemptions for cases involving state secrets or personal privacy, juvenile criminal cases, disputes concluded through mediation, and other documents deemed "inappropriate" (不宜) to publicize (SPC 2013). Revisions to the rules in 2016 expanded the range of publicly available court documents and also began requiring courts to release the case number of any decision deemed unsuitable for posting online, along with an explanation of why the judgment was held back (SPC 2016a).[9] At the same

8. An earlier SPC notice, issued in 2010, had taken a permissive approach, stating that courts could post cases online subject to certain exceptions (SPC 2010).

9. Publicly available documents 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 the termination of a case should be made public unless it

time, the 2016 rules reiterate the principle that local courts have discretion not to post decisions that the courts deem inappropriate for online posting, and they expand the list of case types shielded from public view.[10] 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.[11] To be sure, these carve-outs restrict our view of significant areas of everyday adjudication, such as family law, as well as of how courts resolve difficult or sensitive disputes. And yet even this modest form of controlled transparency signifies a sea change for China's courts.[12]

Why did the court leadership do it? Another angle on this question is political, and the large-scale release of court documents may be viewed first and foremost as a way to serve Party goals by curbing wrongdoing in the courts. Court officials in Henan made this line of argument explicit: judges are more likely to follow the law and less likely to engage in malfeasance when they know their work will be made public. An SPC white paper endorsed this logic in 2017, noting that placing cases online fits with President Xi Jinping's calls for judicial openness and increased public supervision (SPC 2017). Viewed from this perspective, disclosure of court documents was a way for the courts to participate in Xi's signature anticorruption campaign and also join a broader Party-state move toward transparency, a word that entered the Chinese lexicon of governance over the past decade. Freedom-of-information regulations passed in 2007 coexist with a media spotlight glaring enough to expose at least some malfeasance. Officials are also required to release certain types of data, including selected environmental statistics and now court decisions. Somewhat unexpectedly, given the association between democracy and transparency, the contemporary Chinese Communist Party has emerged as a champion of Justice Brandeis's well-known idea that publicity can act as "a remedy for social and industrial diseases" and that "sunshine is . . . the best disinfectant" (1913, 10).

Greater judicial disclosure also appealed to strategic thinkers in the court bureaucracy who thought it could improve the standing of courts with the public or at least help them more effectively monitor and run a far-flung court system. Chinese courts have long been regarded as weak actors, and some thought greater transparency could help improve trust

falls into a specific excluded category. No public posting of the case number or explanation of the reason for nonposting 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. As of early 2019, however, few courts were publicly releasing the case numbers of any nondisclosed cases.

10. This includes cases involving state secrets, crimes committed by minors, divorce cases, cases involving custody or guardianship of children, and most disputes resolved through mediation.

11. The 2016 rules do, however, require that first-instance decisions be made public alongside the appeal when the appellate decision is made public. Before the issuance of the 2016 SPC rules, there was some debate over whether nonfinal 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.

12. Courts have also begun to put large numbers of video recordings of cases online and in 2018 announced plans to place transcripts of court hearings online.

in courts and make it easier for them to resist external pressure.[13] Placing decisions online was also plainly a way to expose the activities of individual judges to their superiors in the court hierarchy. Databases of legal decisions can be a powerful tool of court administration insofar as they alert higher-ups to unusual patterns of decision making, perhaps even identifying individual judges who work exceptionally slowly or routinely hand down outlier decisions.

Technophiles inside China's political-legal system have further argued that algorithms derived from mass digitization of court opinions can advance the efficiency and standardization of 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 reduce judges' workload by drafting parts of opinions or deciding easy cases. A number of experiments are underway to use technology to ensure standardization. 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 (Jie 2016). 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. China's first forays with computer-assisted judging have been small-scale, and some Chinese judges express skepticism about computerized adjudication. Nevertheless, these experiments place new importance on court opinions as the source of data that programmers are now using to construct algorithms used to standardize outcomes or decide cases.

III. DATA AND METHOD: THE HENAN DATA SET

The Henan data set is a collection of 1,058,986 court documents from the Henan High People's Court website (table 1). We chose to focus on cases from a single province, Henan, for three reasons. First, Henan started putting cases online earlier than most other provinces and posted hundreds of thousands of cases before the launch of the SPC website. Studying Henan allows us to look further back in time than would be possible in other provinces. Second, Henan ranks in the bottom third of Chinese provinces in GDP per capita. 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—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 and explore differences among them.

13. 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 the 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 You (2013).

Table 1. Documents in the Henan Data Set

Year	Civil Decisions	Criminal Decisions	Administrative Decisions	Enforcement Decisions	Other	Total
2015	148,023	39,609	7,913	26,134	1,228	222,907
2014	194,616	59,609	9,067	20,926	1,883	286,101
2013	120,050	46,226	3,884	4,240	666	175,066
2012	66,316	30,578	2,429	972	201	100,496
2011	59,379	29,751	2,513	417	222	92,282
2010	60,713	31,331	3,187	428	349	96,008
2009	42,583	17,779	2,566	407	3,363	66,698
2008	1,836	347	142	21	434	2,780
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	15,687	16,378
Total	693,747	255,255	31,710	54,192	24,082	1,058,986

Creating this collection of court documents took 18 months of effort by a team of research assistants with a background in computer science. They developed software to scrape more than 1 million court documents and wrote code known as a "parsing script" to help the computer differentiate between parts of court documents. Owing to the number of documents, a parsing script is necessary for such basic tasks as counting document types or determining how many documents each court uploaded. In terms of format, court opinions follow guidelines set by the SPC, with some local variation.[14] 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, usually 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 yielded table 1, a detailed look at the types of documents contained in the Henan data set.[15] The 1,058,986 court documents span the years between 1996 and 2015, with the great majority dating from 2008–15. Major categories include 693,747 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).

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

15. A few notes on classification: (1) joint civil/criminal cases (刑事附带民事) are classified as criminal, and (2) the "unspecified" row contains 16,378 documents for which the date was missing or obviously wrong (e.g., the year was 2022).

IV. WHAT IS MISSING? ASSESSING BIAS

For all that is available, the Henan data set is also clearly incomplete, and documenting what is missing is an urgent task for researchers. In what follows, 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 does not fully fit with the conventional explanations, such as mediation rates or GDP per capita.[16] In contrast, our analysis suggests that variation in transparency across courts is most likely due to variation in how courts have interpreted the national rules surrounding disclosure. Some courts are more diligent than others in disclosing decisions, especially as court leaders likely vary in how much emphasis they place on transparency.

In 2014, internal court statistics report that Henan courts completed 685,890 cases, compared to 286,101 documents in our collection.[17] On average, this means that Henan courts placed just over 41% 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% of completed cases, compared to just 14% for the least compliant court.[18] This variation is particularly striking for a province that had released rules concerning what must go online by 2014 and had already begun ranking courts based on their compliance with provincial transparency policy.

The variation in disclosure rates across courts has not disappeared. During our research trip in May 2017, many Henan judges and scholars argued that 2014 was early in the adoption of mandatory disclosure and that disclosure rates had dramatically improved and become more uniform since then.[19] There are now clearer SPC rules about what should be disclosed, and disclosure is now the default. In addition, the Henan courts introduced software in mid-2016 that automatically uploads court documents to the SPC

16. Using the province as the unit of analysis, Ma, Yu, and He (2016, 208–9) find a correlation between court disclosure rates and GDP per capita. To the best of our knowledge, no existing research examines whether mediation rates are correlated with court disclosure rates. A commonsense explanation, however, would be that courts with higher mediation rates have lower disclosure rates because there is no obligation to release mediation agreements.

17. The data on court totals are from internal court statistics that were provided by a contact inside the Chinese court system.

18. All estimates should be treated as upper bounds. As discussed later on, courts sometimes publicly post more than one decision related to a case.

19. In May 2017, we presented a draft version of this article at Zhengzhou University and solicited feedback from a mixed audience of scholars, judges, and other legal professionals. A common reaction was that 2014 was ancient history and that the missingness problem should be much improved and possibly totally solved. One judge commented that if we reran our plot of cases for 2016, we would find little variation, with courts all disclosing 80%–90% of their docket. Although the trend is certainly toward greater disclosure of cases, as shown in fig. 1, our analysis shows that the public record is still far from complete.

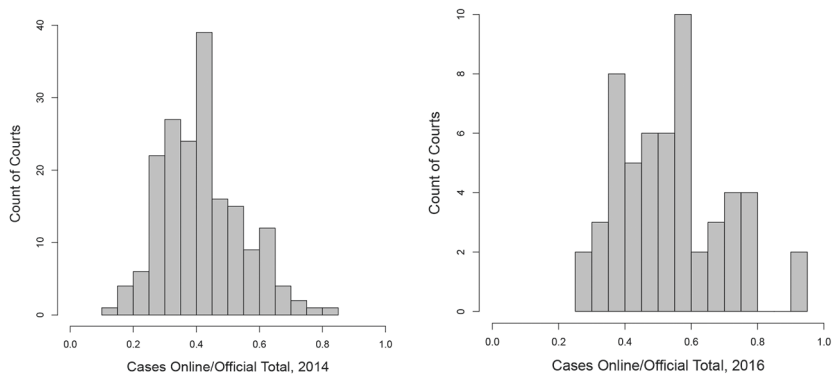

Figure 1. Proportion of cases online by court in the Henan data set, 2014 and 2016

server unless the presiding judge indicates a reason the case should not be made public and a higher-ranking judge in the court approves a request not to disclose. Yet our analysis of 55 courts with publicly available work reports for 2016 shows an average disclosure rate of 54%—an improvement over 2014's average disclosure rate of 41% but hardly the end of the missingness problem (fig. 1).[20] Variation across courts also persists. The highest-ranking court in our 2016 sample has a disclosure rate of 93%, compared to just 26% for the least compliant court.

What might explain the variation in disclosure rates across courts? One obvious explanation is that the national disclosure rules may affect courts differently. For example, mediation agreements are not required to be posted online, and courts with lower disclosure rates could rely more heavily on mediation. However, further analysis of the 2014 data shows that accounting for mediation does not explain differences across courts. In Henan, internal court statistics count 110,134 mediated cases in 2014, and accounting for mediation boosts the average disclosure rate to 52%.[21] Adjusting for mediation also results

20. We were unable to obtain internal court statistics on the total number of cases resolved by each court for 2016. Instead, we looked online for each court's public 2016 court work report, finding work reports—and thus data—for 55 courts. A direct comparison of the same 55 courts in 2014 and 2016 shows a modest jump in disclosure rates from 42% to 54%. Our measure of cases available for 2016 is based on a separate data set of Henan cases posted to the central SPC website, as the Henan High People's Court website stopped posting new cases at the end of 2015. The primary data set used in this article, which was obtained from the Henan High People's Court website, remains a valuable window onto court practices in Henan between 2012 and 2015, particularly because not all of the documents related to those cases migrated to the central SPC website.

21. 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. Henan established one new court in 2016, for a total of 184 courts as of early 2017. To account for mediation, we subtracted the number of mediated cases from the total number of cases the court reported handling. 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 data set.

in even more variation across courts. Although the least compliant court released only enough documents to plausibly cover 20% of completed cases, adjusting for mediation yields four courts that released enough documents to potentially cover 100% of decided cases.[22]

Another obvious place to look for explanations for variation in transparency is court resources. Resource bias suggests that variation in court transparency stems from underlying resource constraints, particularly the availability of personnel to collect judicial decisions, black out personal information, and place them online (Grimmer, Roberts, and Stewart 2019). Nationwide, Ma et al. (2016, 208–9) show a statistically significant correlation between GDP per capita and the 2014 interprovincial disclosure rate, which provides some support for the idea that resource bias might matter. However, our parallel efforts to investigate the relationship among GDP per capita, population, and the 2014 disclosure rates for individual courts within Henan Province show no statistically significant relationship among the variables (app. B; apps. A–C are available online).[23]

Rather, we find evidence within the Henan data that how courts interpret the rules explains at least some of the variation in transparency. In other words, diligence bias—or the degree to which courts scrupulously adhere to national guidelines—helps explain missingness. Divorce cases provide a good example of variation in how closely courts adhered to national rules on disclosure.[24] From January 1, 2014, through September 20, 2016, the SPC rules were that divorce decisions should be made public only if the names of the parties were redacted. Then, in July 2016, the SPC changed the rules to protect personal privacy more fully and prohibited courts from disclosing divorce decisions.[25] In our data set, however, we find that unredacted divorce cases are common before 2016 and that some courts continued posting divorce cases even after the prohibition on releasing them took effect. Our data set contains 29,982 divorce decisions in 2014, an average of 5% of the cases posted in each court, and this average falls to 3.5% of the cases posted in each court only in 2016 (for a total of 38,585 cases posted). As shown in figure 2, in both 2014 and 2016, the proportion of divorce cases posted online is positively correlated with the proportion of the total docket online. This suggests that courts with high transparency rates might be releasing more documents than SPC rules suggest and that courts with low transparency rates are more strictly following the SPC rules.

22. The most compliant court, the Sheqi County basic-level court, released 9% more documents than the number of cases it reported completing in 2014. Courts may release multiple decisions in the same case in order to resolve procedural or jurisdictional issues before issuing a final judgment.

23. We use population as a rough proxy for the business of individual courts, making the assumption that courts located in populous areas are busier than those located in less inhabited areas. During fieldwork in Henan in 2017 we heard from judges that high-population districts might have lower disclosure rates owing to larger per-judge caseloads. If anything, we find a negative relationship between GDP per capita and transparency and a positive relationship between population and transparency, both of which run counter to our intuition.

24. Divorce cases are defined here as civil cases with "divorce" (离婚) in the title of the case.

25. The new rules became effective on October 1, 2016.

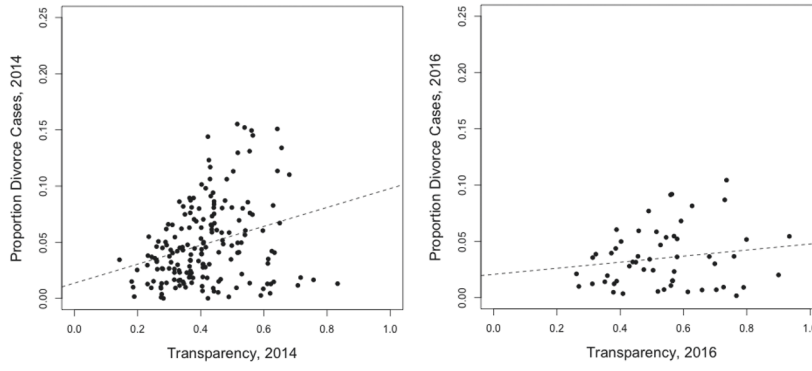

Figure 2. Relationship between transparency and proportion of divorce cases, 2014 and 2016. Both years show a strong correlation between transparency and divorce cases, indicating that court policy on releasing divorce cases may explain variation in transparency.

Diligence bias could also help explain differences in disclosure rates between basic and intermediate courts. Under SPC rules, first-instance decisions may be released only after any subsequent appeal concludes. This rule creates a time lag between first-instance decisions and the posting of those cases online, which may lead to an unintended disclosure gap between basic and intermediate courts. Even though almost all 2014 cases should have been resolved by the time we collected the data, the average intermediate court released enough documents to cover 56% of cases decided in 2014 compared to an average of 40% for basic-level courts, a statistically significant difference ($p < .001$).[26] In 2016, this trend held, with intermediate courts disclosing 63% of documents compared to 51% for basic-level courts, still a statistically significant difference ($p = .047$). Our interpretation is that some first-instance cases may not be posted online because some lower-court judges are unaware when appeals are resolved.[27] Diligence bias also helps explain why there is no relationship between court transparency and GDP per capita. Although our initial expectation might be that resource-rich courts have the human bandwidth to place more documents online, they might also use that capacity to follow SPC standards surrounding nondisclosure strictly, such that fewer documents end up online.

Of course, it is impossible to determine exactly what is not in our data set, and it is also likely that more than one type of bias is at play. For their part, Henan legal professionals

26. Adjusting for mediation rates, intermediate courts released enough documents to potentially cover 70% of all cases decided, compared to an average of 50% for basic-level courts, a statistically significant difference ($p = .0002$).

27. This was a particularly important issue in 2014, when the rules governing the disclosure of first-instance cases were unclear. It is likely less of an issue today as the 2016 rules clarify that first-instance cases should be automatically made public after appeals are decided. In 2014, courts also had widely different approaches to posting the short judgments known as *caidingshu* (裁定书), many of which resolve cases without deciding the merits of a claim or are based on summary procedures, and to disclosing enforcement-related documents.

typically reached for explanations about the importance of court leadership when we asked about variation in disclosure rates during fieldwork there in 2017. Their explanation is that court presidents vary in the value they place on transparency, particularly how much they care about their court's performance in annual province-wide disclosure rankings and the degree to which they prioritize disclosure rates in evaluating individual judges.[28] There is also some evidence that this might be true and that court presidents respond to the incentives created by rankings and evaluation. Ma et al. (2016, 204) 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), and that uploads of judicial decisions to the SPC website spike at the end of the quarter just ahead of court evaluations.

In contrast, American audiences typically speculate that records of politically sensitive cases are suppressed by the court leadership. Again, there is some evidence that this is true. In Henan in particular, lawyers routinely note that cases they have handled involving politically well-connected parties are frequently missing from online databases. Administrative censorship is hard to document, however, because courts are reluctant to discuss it. Technically, Chinese courts are supposed to publicly release a written reason for any judicial decision they choose not to disclose publicly. Even though courts rarely comply with this directive, our analysis of 2,472 such documents includes only three examples of courts willing to either cite the catchall "inappropriate to publish" provision of the rules or admit that they are suppressing collective cases (群体性案件). Of these courts, only a Shandong court disclosed enough detail to identify the type of case held back from public view, a string of collective labor disputes.

In sum, there are three credible explanations for what is missing: administrative censorship, incentive bias, and diligence bias. These three explanations collectively lay the groundwork for future research. What is needed is a sustained, cooperative effort to document the relative importance of different types of bias in China's new online case databases and to trace changes over time.

What to do in the meantime, then, as missing cases will affect every analysis? Building on Grimmer and Stewart's (2013) four principles of automated content analysis, we offer three concrete suggestions about how to learn from a rich, incomplete data set. First, we can seek out pockets of excellent data. Although the results will not be generalizable to all courts, they will provide more certain answers. At the very least, some researchers may decide that a smaller-scale analysis of a more complete corpus is a worthwhile trade-off.[29]

---

28. Specific evaluation metrics for judges and courts are often set by intermediate courts, so the weight given to disclosure rates in court evaluations can vary within a province. Courts that prioritize transparency may strategically place additional documents online to boost their standing in the disclosure rankings. For example, some Henan courts continued to release divorce decisions even after the SPC deemed divorce-related documents unsuitable for online publication. Of course, this behavior also could be due to administrative convenience or clerical errors.

29. Stern (2014) analyzes the caseload of just one court with high disclosure rates and is an example of this approach.

Second, we can use what we know about missingness to bound our estimates of the prevalence of a phenomenon. Frequency estimates of any type of case can be bounded by assuming that all missing cases belong to the set (the upper bound) and then assuming that no missing cases belong to the set (the lower bound). And, third, computational text analysis can be combined with qualitative research. Computational text analysis is particularly valuable preceding fieldwork, when it can be used to uncover patterns and suggest questions that we would not have otherwise known to ask. We are especially optimistic about what we can learn by asking local legal professionals—especially experts or repeat players in a particular area of law—to comment on preliminary results from computational text analysis. There is no substitute for local knowledge, and given the extent of the missing data, checking whether findings are credible is an important reality check. Conversations with Chinese scholars, judges, lawyers, and even plaintiffs can also help suggest explanations for trends, as well as further research questions. We illustrate these three approaches in the next section, using administrative litigation as a case study to illustrate how to work with a large, incomplete data set.

V. COURT DECISIONS AS DATA: INSIGHTS INTO ADMINISTRATIVE LITIGATION

Administrative law is 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, in which citizens use law to challenge illegal state actions. Because of this, administrative litigation lawsuits are often framed as part of the basic repertoire of contention and a way for ordinary Chinese citizens to contest government decisions over sensitive issues such as land seizures, fines, and police detention (Mahboubi 2014; Cui 2017). Both the Chinese- and English-language scholarship have closely tracked the total number of administrative lawsuits, a number that has been treated as a signal of citizen willingness to challenge the state, evidence that the populace is undergoing a "legal awakening" (Zhang and Ginsburg 2019, 285), and an indicator that the Chinese state is increasingly constrained by law (Zhang and Ginsburg 2019).[30]

Despite the fact that the raw number of administrative lawsuits is seen as an important data point for understanding trends in state-society relations, no one has yet looked behind the numbers at what kinds of disputes end up counted in official statistics. In large part, this is because a more granular look became possible only following the Chinese court system's decision to release publicly most court decisions. In what follows, we use a topic model to explore the 25,921 decisions in administrative litigation cases in our data set,

30. For example, a leading Chinese scholar of administrative law writes that "without administrative litigation, many of the plaintiffs would have been still running on the road to petition, and many of the officials would not have heard of terms such as 'excess of power' or 'due process'" (He 2018, 141). Along similar lines, Li's examination of what determines variation in administrative litigation across provinces frames administrative litigation as "the first time in Chinese history [that] victims of official malfeasances may 'routinely' sue the state for remedy" (2013, 815).

with the goal of better understanding the range of cases that collectively constitute "administrative litigation."[31] 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 models to analyze judicial decisions in other countries.[32]

Methodologically, this section shows how topic modeling can lead to new discoveries and, in particular, complicate categories that were previously taken for granted.[33] After sketching a portrait of the data set using topic modeling, close reading of cases revealed a large number of topics related to disputes involving third parties. In these cases, the litigant is using administrative litigation to try to obtain assistance from the state in a dispute with another private individual or organization. We then go on to estimate that 47% of our administrative corpus consists of third-party cases, based on hand-coding a random sample of cases. This finding suggests that a large proportion of administrative lawsuits are private disputes in which litigants are trying to leverage the power and authority of state agencies, rather than efforts to challenge or constrain officials. Mindful of the possibility that missing data could affect our analysis, we then further examine courts with high transparency, bound our estimate, and engage in qualitative fieldwork. All these strategies confirm the prevalence of third-party disputes—a finding that ought to shift the field's interpretation of what a growing administrative caseload means and temper conclusions that it signifies either popular dissatisfaction with the state or a step toward the expansion of judicial review.

Before walking through our research process and results, a brief introduction to topic modeling is helpful.[34] Topic modeling, a tool that originated in computer science and is now used across the social sciences, helped us examine the broad categories of more than 20,000 administrative litigation cases without reading each one. We used the Structural Topic Model package in R (Roberts et al. 2014; Roberts, Stewart, and Airoldi 2016; Roberts, Stewart, and Tingley 2019) to estimate topics, which are groups of words that are likely to appear together within documents. For example, the model suggests that the words "land," "use permit," "collective," "issue," "land used," "construction," and "use" frequently

31. As table 1 shows, the Henan data set includes 31,710 documents from administrative divisions in Henan courts. These include (1) administrative litigation, which comprises lawsuits by individuals or legal persons against the state challenging concrete administrative actions, and (2) nonlitigation enforcement decisions, which are actions brought by administrative entities asking courts to enforce their decisions (usually unpaid fines). We removed 5,789 nonlitigation enforcement cases before running our topic model in order to focus on suits against the state.

32. Our approach also contrasts with recent empirical work on administrative litigation (Zhang, Ortolano, and Lu 2010; Li 2014; Cui 2017) that draws on interviews and small-*n* samples of court documents.

33. Of course, missing data potentially have a large effect on topic models, and the presence of so much missing data in this data set introduces uncertainty to the topic model. In what follows, we argue that our finding about the importance of third-party disputes would hold even accounting for missing cases, but substantial uncertainty remains about how frequently these kinds of disputes occur.

34. For a deeper introduction to topic modeling, see Grimmer, Roberts, and Stewart (2019).

appear together, collectively forming a topic we labeled as "rural land-use permitting."[35] In addition, the model estimates topic proportions for each document, with each document containing a mixture of topics. Research assistants helped us review the highest-frequency words associated with each topic, as well as 20 cases associated with each topic, in order to assign each topic a topic label (e.g., "birth planning" or "withdrawals—individuals").[36] A list of all 82 topics and the highest-frequency words associated with them appears in appendix A.

Six broad categories of disputes emerged from the topic model—topics related to land and property; procedure; fines or punishment; benefits, labor rights, or compensation; general words; and permits and registration. We assigned each topic to one or two of these categories, based on the most important themes in the 20 cases we reviewed (table 2). To describe the range of topics and categories, figure 3 is a correlation plot coded along these six themes, in which the node size is proportional to the amount the topic is discussed within the corpus of text.[37] In other words, larger nodes indicate more common topics. Edges, or the lines connecting nodes, indicate that there is a correlation between topics greater than 0.01 or that the topics are more likely to appear within the same document. In addition, groups of topics that are likely to appear together within documents also come into focus. For example, a group of topics related to the benefits, labor, and worker compensation theme cluster in the right side of the plot. These documents are likely to contain similar language related to monetary demands, particularly related to health and retirement benefits. Land and property claims largely congregate toward the bottom left of the plot. Language related to court procedure and fines appears in many types of cases, and such topics are scattered throughout the correlation plot.

35. A significant amount of preprocessing of the text is required before running a topic model. In brief, we segmented the Chinese text into words using the Stanford Natural Language Processing Chinese segmenter. By examining word removal lists, we found that removing words that appeared in more than half of the documents removed common legal terms that were unlikely to distinguish between cases, so we added this as a preprocessing step. We estimate the number of topics from the data using an algorithm developed by Lee and Mimno (2014), implemented in the STM package in R (Roberts et al. 2019). We chose to use the Lee and Mimno algorithm to choose the number of topics because our goal was simply to describe the data, and this imposed an external criterion on this decision. We also found that the Lee and Mimno algorithm provided topics with high interpretability; a close reading of topics after the model was fit indicated that they had clearly separable ideas. However, we note that preprocessing decisions can have an important impact on results (Denny and Spirling 2018) and that there are likely many other ways of categorizing the same data.

36. As Grimmer and Stewart (2013) note, close reading is essential to ensure that topic labels accurately capture the content of the documents. The 20 cases we read included 10 cases that the model identified as most representative of the topic—i.e., the 10 cases with the highest percentage of words correlated with the topic. We also reviewed 10 additional randomly selected cases with an estimated topic proportion above 0.3, meaning that at least 30% of the words in the document were estimated to be from the topic. We did this to make sure the topic label corresponded to the variety of cases with high proportions of the topic and to ensure that the topic was not skewed by large numbers of nearly identical cases.

37. In some cases, we grouped topics into more than one category. In such cases, we identify both categories in app. A but assign the topic the predominant category in fig. 3.

Table 2. Categories of Administrative Litigation in the Henan Data Set

Category	Proportion of Topic Model (%)
Land and property disputes only	30.7
Topics related to procedural decisions only	27.8
Disputes over fines and punishments assessed by administrative agencies only	11.1
Disputes over benefits, labor rights, or compensation for workplace injuries only	10.6
Topics that include land and property as well as procedure	8.8
General word topics, with litigation-related words common to many kinds of cases*	6.9
Topics relating to permits and registration (but not land or property)	2.0
Topics that include land and property as well as benefits, labor rights, and compensation	1.2
Topics that include fines and punishment as well as benefits, labor rights, and compensation	<1
Topics that include fines and punishment as well as land and property	<1

Note.—We put each topic into one or two broad categories; i.e., some topics were assigned to one broad category and some to multiple categories. We then estimated the proportion of words in each category by estimating the expected topic proportions across documents assigned to that category.

* These are general litigation-related words, such as "respondent" and "rights," that persist even after removing words that appear in more than half of the court cases.

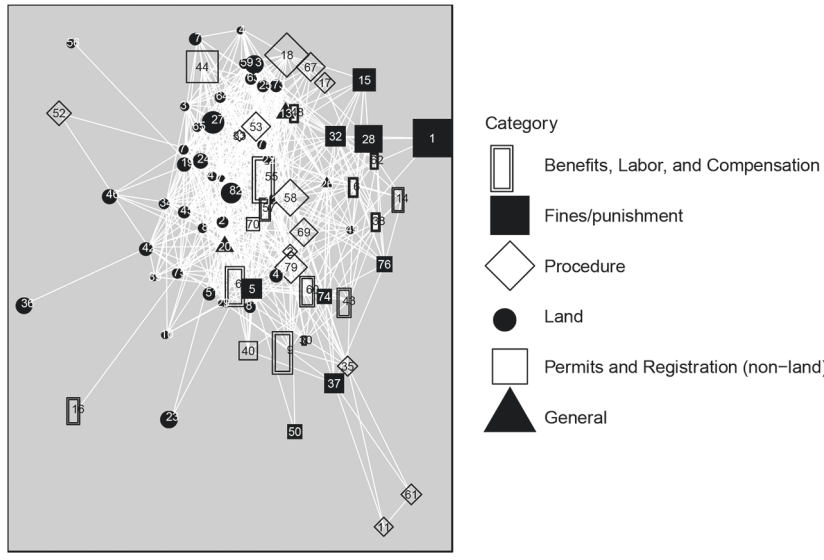

Figure 3. Correlation plot of administrative cases in the Henan data set. Topics 54, 66, 78, and 80 were removed because they were not connected to other topics.

Using the topic model output, we can estimate the proportion of the documents related to each of the broad categories of topics. Table 2 shows the sum of the topic proportions for topics related to each of the broad categories in the topic model. Overall, we see that the topic model estimates that 30% of the corpus is related to land and property disputes. In addition, 11% of the cases are related to fines and punishment.

By closely reading the cases most representative of each topic, we discovered that 37 of 82 topics mark an attempt to draw the state into private disputes, typically by challenging an administrative decision in favor of another individual, group, or entity. Sometimes, the fact that the underlying dispute is between private parties is obvious because a designated third party (第三方) appears in the list of parties that opens all Chinese court decisions. In a 2013 case, for example, a concrete factory sued the Dengfeng Human Resources and Social Security Bureau over a decision to award work-related injury compensation to Wencan Liu. The factory had been privatized 4 years before the lawsuit and argued (unsuccessfully) that they should not bear liability for a case of tuberculosis caused by a 2-decade work history that largely fell into the period when it was a state-owned enterprise.[38] Another example of this type of third-party dispute is a 2014 case from Pingdingshan City involving a complicated property dispute—and attempt to kick out several alleged squatters with a competing claim—in which the Real Estate Administration Bureau was dragged to court to (successfully) defend its property registration decision.[39] When the third party was not clearly identified at the start of the decision, it took more attentive reading to surface the underlying civil dispute. In topic 44, for example, there were a number of cases challenging business registration with the State Administration of Industry and Commerce. The text of these decisions, however, makes clear that these disputes are in fact about the underlying ownership of the company in question.

Having uncovered the importance of third-party disputes through topic modeling, we sought to estimate the frequency of third-party disputes by hand-coding a random sample of 500 cases. This was also an important validation step, along the lines recommended by Grimmer and Stewart (2013), to check findings from an unsupervised topic model. Research assistants with a background in Chinese law decided whether each case represented a dispute between two private parties or between a private party and the government.[40]

38. Dengfeng Shaolin Cement Company v. Dengfeng Human Resources and Social Security Bureau (2013), Zhengzhou Intermediate Court. Court decision on file with the authors.

39. XX Jin and YY Liu [names anonymized by the court] v. Pingdingshan City Real Estate Administration Bureau (2014), Pingdingshan Xinhua District People's Court. Court decision on file with the authors.

40. Coding rules that lay out our detailed definition of a "third-party dispute" appear in app. C. Each case was read and coded independently by two research assistants. The coders started with a basic definition of third-party cases and then reviewed an initial 100 cases in order to develop a more precise definition of third-party cases. They then each reviewed the same 500 cases in the random sample. In the 34 instances in which their coding diverged, one of the authors reviewed the case with both coders to arrive at a final classification. The definition of third-party cases that we used is set forth in app. C.

Overall, 47% of cases ($n = 234$) were third-party disputes, whereas 40% ($n = 201$) of cases reflected disputes with the government. (In 13% of cases, we were not able to make a determination based on the text of the decision.) At first glance, these numbers are striking: third-party disputes are approaching a majority of administrative cases in our sample. This is a finding that would surprise even close observers of the Chinese legal system. Although there is a parallel trend in criminal law in which litigants sometimes use criminal sanctions to seek compensation in private disputes owing to difficulty enforcing civil judgments (Liebman 2015), existing work on Chinese administrative law has not been attentive to how administrative litigation can get piggybacked onto civil disputes.

VI. ACCOUNTING FOR MISSING DATA

How do missing data affect our analysis? On one hand, perhaps our data set is missing cases brought against the state owing to their political sensitivity. On the other hand, perhaps our data set contains only a partial view of third-party disputes because courts hold varying interpretations of what they are required to disclose or simply devote insufficient resources to ensuring cases are placed online. In order to buttress our finding and also illustrate how to work with incomplete data, we use three approaches to make sure third-party disputes remain a substantial portion of the docket even accounting for missing data.

As a first step, we look only at courts with well above average transparency to see whether the analysis changes. Within our random sample of 500 cases, 63 were from courts with higher than 65% transparency. In these high-transparency courts, our coding suggests that 38% of cases were third-party cases, 46% of cases represent an underlying dispute with the government, and 16% could not be determined from the text of the decision. Over one-third of administrative lawsuits are still third-party disputes even in high-transparency courts for which missing data are less of a problem.

Second, we draw bounds around our estimate of the prevalence of third-party disputes. From the analysis we described in the previous section, the corpus contains about 41% of all types of cases in 2014. The province-wide disclosure rate for administrative cases was higher than for the corpus as a whole: our data set includes 9,067 administrative cases from 2014, out of a total of 15,453 administrative cases that Henan courts reported resolving in 2014—a disclosure rate of 59% (Henan High People's Court 2015). We therefore use this 59% transparency rate in administrative cases to bound our estimate. Assuming our estimate that 47% of all cases are third-party cases is correct, we can calculate the lower-bound proportion by assuming that all cases that are missing are not third-party cases. This would leave us with 28% of cases being third-party cases, still a substantial number. But if all missing cases were third-party cases, then a full 69% of all cases would be third-party cases. In either case, we believe that such a substantial portion of cases deserves more attention in the literature.

Finally, one of the authors traveled to China in 2018 to present our findings about the prevalence of third-party disputes to audiences that included Chinese judges and

administrative law professors.[41] In cooperation with local universities, we held interactive forums in Beijing, Shanghai, Chengdu, and Changchun that closed with a specific request for feedback about whether listeners had noticed that administrative litigation often masks a civil dispute. Across all four cities, audience participants noted a close fit between our findings and their experiences. In particular, judges noted that third-party disputes are common. "We have a sense of administrativization of civil cases," said a Beijing judge in a typical reaction to our slide deck. Although it is possible that audience members were disposed to politely agree with our findings, we take solace in the fact that at least some participants felt comfortable publicly expressing disagreement. In particular, a few participants pushed back against our interpretation of our findings and disputed whether third-party disputes are common or, as one judge in Beijing put it, only "a sprout." But the majority of commentators agreed that the courts are often caught in the middle between private disputants and that judges have no option but to become involved even if they would prefer not to hear or resolve the case. One judge, expressing his frustration at being drawn into such cases, drew a link to the wider phenomenon of what Chinese judges call "abusive litigation" (滥用诉讼). One scholar noted that courts are taking steps to address the frequency of private litigation being transformed into administrative litigation and predicted that over time our data would begin to show a decline in the percentage of cases involving third-party disputes.

Regardless of the exact number of third-party cases, their prevalence in our corpus highlights the limits of conceptualizing administrative litigation primarily as a tool for angry people to petition a repressive or unresponsive state. To be sure, some cases certainly fit this mold. The topic model includes topics challenging reeducation-through-labor decisions (topic 1), appeals relating to petitioning (topic 2), claims by individuals detained for petitioning to Beijing (topic 28), and challenges to birth-planning fines (topic 50). However, the "citizens versus the state" frame flattens away a huge amount of complexity. Looking closely at our corpus makes plain that the aggregate number of administrative cases is a poor measure of citizen willingness to sue the state, even though observers of Chinese administrative law often treat it this way. In addition to third-party disputes, the prevalence of case strings shows how easy it is to count the same dispute numerous times. The huge number of documents related to court procedure—which are generally counted in official court tallies of administrative litigation—also suggests that the aggregate number of administrative lawsuits includes many mundane procedural decisions,

41. We also presented initial findings to Chinese audiences in 2017, including in Henan, but without the same emphasis on the prevalence of third-party disputes. All discussions were hosted at local universities or courts, typically with Chinese faculty serving as interlocutors to help ensure attendance from legal professionals with administrative law practice experience and expertise. Both the presentation and subsequent discussion were conducted in Chinese and lasted between 90 and 120 minutes. The presenting author took detailed notes on the discussions that followed the presentation, and all quotes in this paragraph are drawn from those notes.

such as remands for new trials (topics 17 and 18) and decisions related to statutes of limitations (topic 53) and jurisdiction (topic 52).

All of this shows how topic modeling can be a useful tool of discovery to surface underappreciated motifs in administrative litigation, even when working with incomplete data. The prevalence of third-party disputes, in particular, is not a theme that would necessarily emerge from examining national statistics, tracking the news, or reading scholarship. Rather, it reflects the vantage point of the data: 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 Chinese administrative judges actually spend their time and reminds us how much of administrative law is about leveraging state authority rather than challenging it.

VII. CONCLUSION

A new wave of digital scholarship on law and the courts is already underway, as researchers start to bring the tools of computational text analysis to bear on the texts core to legal scholarship. For many of us, this digital turn is exciting because of the ways growing digital archives can contribute to knowledge across disciplines. As collections of court decisions grow, for example, legal scholars will be able to trace evolving interpretations of concepts such as fault, causation, or damages or to examine patterns in how courts handle specific types of cases or parties. Political scientists can also 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. At the same time, however, the millions of Chinese court decisions now online will not be the only corpus that is tantalizingly large and frustratingly incomplete. Digitizing documents offers an opportunity for censorship, and there is evidence that this happens. Selected legal academic articles from the 1950s, for example, no longer appear in China's two leading online databases (Tiffert 2019). And even when censorship does not occur, commercial databases have incentives to oversell the completeness of their product rather than account for what is missing. Our general call is for legal scholars to take the missingness problem seriously and consider following one of the three strategies detailed above, rather than succumbing to the temptation to treat even a very large-*n* sample as an accurate reflection of reality.

At the same time, taking missing data seriously does not mean giving up. Our investigation of administrative litigation shows how topic modeling opens a wide-angle perspective on courts' daily activity, exposing unappreciated trends and also surfacing new questions for research. Answering the deeper "why" and "how" questions, though, will likely continue to require the type of information about local context that typically emerges from time on the ground, especially as court judgments provide only one, often limited, view of actual practice. Thus, we hope that the migration toward treating text as data in the field of law will also mean a surge of multimethod work that combines digital tools with interviews, participant observation, archival work, and close reading. After all, "computers

amplify human abilities," not replace them (Grimmer and Stewart 2013, 4). This also means that the best work is done in teams that combine skills and insights from law, the social sciences, and computer science.

Theoretically, the sudden availability of so many data from China's courts raises questions that our field will grapple with for years to come. Overall, 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 is no longer obtaining data but rather effectively using the public record and prioritizing among research questions.

REFERENCES

Brandeis, Louis D. 1913. "What Publicity Can Do." *Harper's Weekly*, December 20, 10–13.

Bryan, Amanda C., and Eve M. Ringsmuth. 2016. "Jeremiad or Weapon of Words? The Power of Emotive Language in Supreme Court Dissents." *Journal of Law and Courts* 4 (1): 159–85.

Cai, Lidong, and Siming Liu. 2016. 社会团体法人自治与司法审查的实证研究 [An empirical study on the autonomy of social enterprise legal persons and judicial review]. *Faxue Zazhi* 法学杂志 12:15–22.

Cui, Wei. 2017. "Does Judicial Independence Matter? A Study of the Determinants of Administrative Litigation in an Authoritarian Regime." *University of Pennsylvania Journal of International Law* 38:941–98.

Denny, Matthew J., and Arthur Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It." *Political Analysis* 26 (2): 168–89.

Evans, Michael, Wayne McIntosh, Jimmy Lin, and Cynthia Cates. 2007. "Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research." *Journal of Empirical Legal Research* 4 (4): 1007–39.

Finder, Susan. 1989. "Like Throwing an Egg against a Stone? Administrative Litigation in the People's Republic of China." *Journal of Chinese Law* 3:1–10.

Funk, Kellen, and Lincoln A. Mullen. 2018. "The Spine of American Law: Digital Text Analysis and U.S. Legal Practice." *American Historical Review* 123 (1): 132–64.

Grendstad, Gunnar, William R. Shaffer, and Eric N. Waltenburg. 2017. "Norway: Managed Openness and Transparency." In *Justices and Journalists: The Global Perspective*, ed. Richard Davies and David Taras, 235–54. New York: Cambridge University Press.

Grimmer, Justin, Margaret E. Roberts, and Brandon Stewart. 2019. "Text as Data: How to Make Large Scale Inferences from Language." Unpublished manuscript.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 1–31.

He, Haibo. 2018. "How Much Progress Can Legislation Bring: The 2014 Amendment of the Administrative Litigation Law of PRC." *University of Pennsylvania Asian Law Review* 13:137–90.

He, Weifang. 2003. 建设透明法院 [Creating transparent courts]. *Southern Weekend*, April 2. https://perma.cc/ZD5Z-HXL7.

Henan High People's Court. 2015. 河南省高级人民法院工作报告解读河南省第十二届人民代表大会第四次会议 [Interpreting the fourth meeting of the twelfth people's congress of Henan Province—the Henan Province High People's Court work report]. https://perma.cc/KNK4-KSQQ.

Hollyer, James R., B. Peter Rosendorff, and James Raymond Vreeland. 2015. "Transparency, Protest, and Autocratic Instability." *American Political Science Review* 109 (4): 764–84.

Ingram, Mathew C. 2017. "Uncommon Transparency: The Supreme Court, Media Relations, and Public Opinion in Brazil." In *Justices and Journalists: The Global Perspective*, ed. Richard Davies and David Taras, 58–80. New York: Cambridge University Press.

Jie, Zhu. 2016. 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 for the people and fair justice]. Supreme People's Court. https://perma.cc/6M9T-7QE4.

Klingenstein, Sara, Tim Hitchcock, and Simon DeDeo. 2014. "The Civilizing Process in London's Old Bailey." *Proceedings of the National Academy of Sciences of the USA* 111 (26): 9419–24.

Law, David S. 2016. "Constitutional Archetypes." *Texas Law Review* 95:153–243.

———. 2019. "Constitutional Dialects and Transnational Legal Orders." In *Constitution-Making and Transnational Legal Order*, ed. Gregory Shafer, Tom Ginsburg, and Terence C. Halliday, 110–55. New York: Cambridge University Press.

Lee, Moontae, and David Mimno. 2014. "Low-Dimensional Embeddings for Interpretable Anchor-Based Topic Inference." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1319–28. Red Hook, NY: Curran.

Li, Ji. 2013. "Suing the Leviathan—an Empirical Analysis of the Changing Role of Administrative Litigation in China." *Journal of Empirical Legal Studies* 10 (4): 815–46.

———. 2014. "Dare You Sue the Tax Collector? An Empirical Study of Administrative Lawsuits against Tax Agencies in China." *Pacific Rim Law and Policy Journal* 23 (1): 57–112.

Liebman, Benjamin L. 2011. "The Media and the Courts: Towards Competitive Supervision?" *China Quarterly* 208:833–50.

———. 2015. "Leniency in Chinese Criminal Law? Everyday Justice in Henan." *Berkeley Journal of International Law* 33 (1): 153–222.

Liebman, Benjamin, and Tim Wu. 2007. "China's Network Justice." *Chicago Journal of International Law* 8:257–321.

Livermore, Michael A., Allen B. Riddell, and Daniel N. Rockmore. 2017. "The Supreme Court and the Judicial Genre." *Arizona Law Review* 59:837–902.

Lorentzen, Peter, Pierre Landry, and John Yasuda. 2014. "Undermining Authoritarian Innovation: The Power of China's Industrial Giants." *Journal of Politics* 76 (1): 182–94.

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): 254–77.

Ma, Chao, Xiaohong Yu, and Haibo He. 2016. 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]. *China Law Review* 中国法律评 4:195–246. https://perma.cc/6T5V-KNXH.

Mahboubi, Neysun. 2014. "Suing the Government in China." In *Democratization in China, Korea and Southeast Asia? Local and National Perspectives*, ed. Kate Xiao Zhou, Shelley Rigger, and Lynn T. White III, 141–55. New York: Routledge.

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): 762–86.

Miller, Ian Matthew. 2013. "Rebellion, Crime and Violence in Qing China, 1722–1911: A Topic Modeling Approach." *Poetics* 41 (6): 626–49.

National People's Congress Standing Committee. 2012. 中华人民共和国民事诉讼法 [Civil Procedure Law of the People's Republic of China]. https://perma.cc/MBX4-V2XX.

Nguyen, Trang Mae. 2019. "In Search of Judicial Legitimacy: Criminal Sentencing in Vietnamese Courts." *Harvard Human Rights Journal* 32 (1): 147–88.

Patton, Dana, and Joseph L. Smith. 2017. "Lawyer, Interrupted: Gender Bias in Oral Arguments at the US Supreme Court." *Journal of Law and Courts* 5 (2): 337–61.

Rice, Douglas R. 2017. "Issue Divisions and US Supreme Court Decision Making." *Journal of Politics* 79 (1): 210–22.

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.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. "STM: R Package for Structural Topic Models." *Journal of Statistical Software* 91 (2): 1–40.

Sina Court Channel. 2016. 中国裁判文书网：全球最大文书公开平台10月1日起施行 [China court judgments: World's largest online judgments platform to come into effect October 1]. https://perma.cc/9KQQ-HQB4.

SPC (Supreme People's Court). 2009a. 人民法院第三个五年改革纲要 [Third 5-year reform outline for the people's courts]. February 24. https://perma.cc/U9V9-5PQ5.

———. 2009b. 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"]. December 8. https://perma.cc/TT2U-J4WB.

———. 2010. "关于人民法院在互联网公布裁判文书的规定" 和 "关于人民法院直播录播庭审活动的规定" 的通知 [Notice of the Supreme People's Court on the issuance of judicial documents on the internet by the people's courts and the provisions on the live broadcasting and rebroadcasting of court trials by the people's courts]. November 21.

———. 2013. 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.

———. 2016a. 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. 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 specifications for civil decisions by the people's courts 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." March 14. https://perma.cc/MN9U-5K6L.

Staton, Jeffrey K. 2010. *Judicial Power and Strategic Communication in Mexico*. New York: Cambridge University Press.

Stern, Rachel E. 2014. "The Political Logic of China's New Environmental Courts." *China Journal* 72:53–74.

Tiffert, Glenn D. 2019. "Peering Down the Memory Hole: Censorship, Digitization, and the Fragility of Our Knowledge Base." *American Historical Review* 124 (2): 550–68.

Wu, Hongyao, and Liang Zhang. 2017. 死刑复核程序中被告人的律师帮助权—基于255份死刑复核刑事裁定书的实证研究 [Defendants' right to assistance from an attorney in the death penalty review process: An empirical study based on 255 criminal verdicts of death penalty review]. *Falü Shiyong* 法律适用 7:61–69.

You, Wei. 2013. 用制度排除权力干预维护司法权威 [Use systems to prevent political intervention and safeguard judicial authority]. *Legal Daily*. https://perma.cc/KBX2-QHDJ.

Zhang, Taisu, and Tom Ginsburg. 2019. "China's Turn toward Law." *Virginia Journal of International Law* 59:278–361.

Zhang, Xuehua, Leonard Ortolano, and Zhongmei Lu. 2010. "Agency Empowerment through the Administrative Litigation Law: Court Enforcement of Pollution Levies in Hubei Province." *China Quarterly* 202:307–26.

Zhao, Ruigang, and Xieyang Geng. 2016. 指导性案例 "适用难" 的实证研究—以261份裁判文书为分析样本 [An empirical study on the "difficulties in applying" guiding cases: Taking 261 judgments as a sample for analysis]. *Faxue Zazhi* 法学杂志 3:115–23.