Peering down the memory hole: history, censorship and the digital turn

Glenn Tiffert[1]

Economic, technological and political forces are fundamentally reshaping how knowledge is produced and consumed. Much as clay tablets and bone gave way to papyrus and parchment, paper is now yielding to bitstreams. In turn, libraries are rebalancing their collections away from physical volumes towards online resources, where a small number of remote providers dominate storage and delivery, and effectively govern the terms of access for all. For many, printed books and periodicals are fast becoming reminders of things past.[2]

This digital turn is seductive; it promises to deliver abundant knowledge with unprecedented convenience and immediacy. Yet, commentators have noted that it also concentrates power, exacerbates inequalities, and raises thorny theoretical, ethical, and methodological concerns.[3] For instance, a handful of media conglomerates now control more than half of all academic publishing, which makes it highly vulnerable to rent-seeking, poor stewardship, or deliberate manipulation. The oligopolistic pricing power these publishers enjoy, protected by copyright, delivers enviable profit margins while squeezing even the richest institutions.[4]

Though the questions are many, censorship has scarcely registered among them. Arguably, this reflects the privileged position of those who dominate the conversation: scholars who study and work in comparatively liberal societies, where one can reasonably assume that a publication accessible today will remain so tomorrow, allowing for normal conservation.

This study dispels that innocence. Using a case from the People's Republic of China (PRC), it explores some of the most disturbing ramifications of our deepening digital dependence. Simply put, the Chinese government is leveraging technology to quietly export its domestic censorship regime abroad and, by manipulating how observers everywhere comprehend its

---

[1] Glenn Tiffert is the Lieberthal-Rogel Center for Chinese Studies Distinguished Postdoctoral Fellow in Residence at the University of Michigan, Ann Arbor. The Lieberthal-Rogel Center for Chinese Studies generously funded this study. At the University of Michigan, Professors Mary Gallagher, Steven Abney, and Kerby Shedden offered valuable feedback. Fu Liangyu, the university's Chinese Studies librarian, helped to locate and acquire materials. Luo Fusheng, Arden Shapiro, Yan Wei, and Yin Shiyuan provided essential research assistance. Margaret Orton coded the naïve Bayes classifier.

[2] Watanabe, Teresa. "Universities Redesign Libraries for the 21st Century: Fewer Books, More Space." *Los Angeles Times* (2017): Accessed July 26, 2017. http://www.latimes.com/local/lanow/la-me-college-libraries-20170419-story.html.

[3] Roy Rosenzweig, "Scarcity or Abundance? Preserving the Past in a Digital Era," *American Historical Review* 108, no. 3 (2003): 735-762; Tim Hitchcock, "Confronting the Digital: Or How Academic History Writing Lost the Plot," *Cultural and Social History* 10, no. 1 (2013): 9-23; Ludmilla Jordanova, "Historical Vision in a Digital Age," *Cultural and Social History* 11, no. 3 (2014): 343-348; Lara Putnam, "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast," *American Historical Review* 121, no. 2 (2016): 377-402. Toni Weller, ed. *History in the Digital Age* (New York: Routledge, 2013).

[4] Vincent Larivière, Stefanie Haustein, and Phillippe Mongeon, "The Oligopoly of Academic Publishers in the Digital Era," *PLOS One* 10, no. 6 (2015): 5.

past, present, and future, it is enlisting them without their consent in an alarming project to sanitize the historical record and globalize its own competing narratives.[5]

The practices exposed here instantiate a protean hazard that confronts us all. But in the pages that follow, I aim to illustrate their gravity in a way that also showcases how technology can enhance our capacity to resist. The study therefore concludes with some reflections on what is to be done, followed by an appendix detailing the tools I used, my workflow, and methodology. Surely, one good digital turn deserves another.

Defining the Problem:

For consumers and producers of knowledge, censorship is a first order problem. It is epistemologically corrosive because it contaminates everything it touches, and even sound research practice may offer no defense. Perversely, the more faithful scholars are to their censored sources, the better they may unwittingly promote the biases and agendas of the censors, and lend those the independent authority of their professional reputations. Historians are taught to mitigate such risks by reading the verisimilitude of their sources critically, but this training tends to treat sources as a stable bequest and frame censorship retrospectively rather than as a dynamic going concern.[6] That must change.

In the past, censors altered history by striking offensive passages, tearing out pages, and seizing or destroying entire texts, all crude methods by today's standards. Now, they can tinker endlessly with the digital record to achieve their goals without ever leaving their desks, making one non-destructive edit after another, each propagating nearly instantaneously around the globe, leaving behind no discernible trace or loose ends. Algorithms can automate the process and catapult it far beyond human limitations. In short, technology has elevated Orwell's grotesque metaphor of history as a palimpsest, "scraped clean and reinscribed exactly as often as was necessary," to a terrifying level of efficiency.[7]

In the PRC, this is already happening. The Chinese government fuses the hegemonic instincts and mobilizational infrastructure of a classic Leninist regime with deep pockets and cutting-edge technologies. Its aggressive policing of social media, art, news, and scholarship sets the bar for like-minded governments everywhere, earning it the sobriquet "The People's Republic of Amnesia."[8] Lately, Chinese leaders at the highest levels have singled out "historical nihilism," shorthand for challenging orthodox narratives about the past, for special attention.

---

[5] He Yiting 何毅亭, "Ershiyi shiji shi Zhongguo huayu fuxing de shiji." 二十一世纪是中国话语复兴的世纪 [The Twenty-First Century is the Century of the Revival of Chinese Discourse] *Xuexi shibao* 学习时报 [*Study Times*] (May 29, 2017): Accessed June 24, 2017. http://www.studytimes.cn/zydx/GCFT/2017-05-29/9532.html.

[6] Arlette Farge, *Le Goût De L'Archive* (Paris: Seuil, 1989).

[7] George Orwell, *Nineteen Eighty-Four* (New York: Alfred A. Knopf, 2009), 42.

[8] Louisa Lim, *The People's Republic of Amnesia: Tiananmen Revisited* (New York: Oxford University Press, 2015). Gary King, Jennifer Pan, and Margaret E. Roberts, "How Censorship in China Allows Government Criticism But Silences Collective Expression," *American Political Science Review* 107, no. 2 (2013): 1-18 ; Gary King, Jennifer Pan,

Those in China who run afoul of the censors confront a sliding scale of penalties, including harassment by the authorities, closure of publications and online accounts, humiliating investigations into personal affairs, business activities and tax status, and ultimately unemployment, eviction, and criminal prosecution. A recent provision of law also exposes "those who infringe upon the name, likeness, reputation, or honor of a hero, martyr, and so forth, harming the societal public interest" to civil liability.[9]

Facing such pressure, previously outspoken intellectuals and activists are going silent, and publishers, for fear of losing access to the Chinese market, are faltering. In 2017, Cambridge University Press (CUP) censored 315 articles and reviews from its catalog of a venerable British academic journal, *The China Quarterly*.[10] The affected content was briefly unavailable from the CUP website in China until CUP reversed itself following international condemnation. Chinese authorities also asked CUP to censor approximately 100 articles from the *Journal of Asian Studies*, the flagship journal of the U.S.-based Association for Asian Studies.

How does this impact us? Ordinarily, scholars expect that when online platforms such as JSTOR or ProQuest carry a publication, those platforms will faithfully reproduce the original and note any material omissions or changes. This frees libraries to clear redundant volumes from their stacks or forego purchasing those volumes altogether, which can trim expenses and open up valuable space for other purposes. Recently, a number of such platforms have appeared in the PRC and, disarmed by familiar-looking veneers, foreign libraries are extending them comparable trust.

This is deeply misguided. As the CUP episode demonstrates, digital platforms in the PRC are subject to the demands of a mercurial domestic censorship regime and the authoritarian government behind it. In purely commercial terms, this precludes them from delivering full replacement value to their subscribers, or guaranteeing that a text accessible today will remain

---

and Margaret E. Roberts, "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument," *American Political Science Review* (forthcoming). Sarah Cook, *The Long Shadow of Chinese Censorship: How the Communist Party's Media Restrictions Affect News Outlets Around the World* (Washington DC: Center for International Media Assistance, National Endowment for Democracy, 2013).

[9] Josh Chin, "China Set to Tweak Civil Code to Punish Revisions of Martyr Lore." *Wall Street Journal: China Real Time Report* (2017): accessed June 17, 2017, http://on.wsj.com/2miOVcF. Chi Fangxu 迟方旭, "'Minfa zongze' di 185 tiao de hexin yaoyi shi weihu shehui gonggong liyi." 《民法总则》第 185 条的核心要义是维护社会公共利益 [Safeguarding the Common Interests of Society is the Core Point of Article 185 of the 'General Principles of Civil Law'] *Qiushi* 求是 [*SeekingTruth*] (June 7, 2017): accessed June 17, 2017, http://www.qstheory.cn/dukan/hqwg/2017-06/07/c_1121099778.htm.

[10] Johnson, Ian. "Cambridge University Press Bends on Censorship, Pulling Articles on Site in China." *New York Times*, August 18, 2017: A7.

so tomorrow, without modification.[11] Yet, tempted by the riches on offer, libraries are increasingly outsourcing critical parts of their Chinese collections to such platforms. Motivated by thrift and efficiency, some are going still further by unilaterally deprecating their physical holdings, an alarming development that not only cedes control of a crucial portion of the source base from which tomorrow's scholarship on China will be written, but also eliminates the independent evidence essential for detecting, exposing, and circumventing the exercise of censorship today and in the future.

Let us make this concrete. When the PRC was established in 1949, the new communist government abrogated the inherited legal system, and proposed to construct a socialist successor free of contamination from the *ancien régime*. In the years that followed, the PRC's two leading academic law journals recorded the sometimes fierce struggles over what this new legal system should look like. *Zhengfa yanjiu* 政法研究 [*Political-Legal Research*] comprised 62 issues between 1954 and 1966, when the Cultural Revolution forced its closure. Sponsored by the Chinese Association for Politics and Law, in Beijing, it counted many of the highest legal officials in the central government among its patrons, and its coverage generally favored their statist priorities. By contrast, *Faxue* 法学 [*Law Science*] enjoyed a much shorter life. It comprised just eighteen issues, all published between 1956 and 1958, when the Anti-Rightist Campaign forced its closure.[12] *Faxue* was sponsored by the East China Institute of Politics and Law in Shanghai, one of a handful of regional academies established in the early 1950s by the PRC state to train a new generation of socialist cadres for administrative and legal positions.

No institution in the United States holds a complete, original print run of *Zhengfa yanjiu*, and *Faxue* is nearly as rare. Some have microfilm or reprint editions, usually stored off-site.[13] For most researchers, online access is the norm. Two of the four leading online platforms in the PRC offer full-text articles from these journals: China National Knowledge Infrastructure 中国知网 [CNKI], a commercial venture connected to Tsinghua University, and the National Social

---

[11] Foreign subscribers to the online edition of the People's Daily 人民日报, for example, may get different results depending on if their vendor hosts its servers in China or not. Some search terms may not generate any results and instead trigger an automatic connection reset that locks the user out of the database for a minute or more.

[12] The first three issues, dating from 1956, bear the title *Huadong zhengfa xuebao* 华东政法学报 [*East China Journal of Politics and Law*]. The publication was renamed *Faxue* beginning with the first issue in 1957.

[13] A five-volume reproduction of *Zhengfa yanjiu* was published in Tokyo in 1967. The issues found in the United States tend to be from this edition. *Zhengfa yanjiu* (Tokyo: Da'an, 1967). A two-volume reprint of the *Faxue* issues examined in this study was published in Shanghai in 2012. Owing to competition from the online edition, very few institutions in the United States own it. When I attempted to buy copies, PRC vendors informed me without explanation that the volumes could not be shipped to my overseas address, though other books on Chinese law in the same orders presented no obstacle. I obtained them through other channels and discovered that the reprints are very good but not perfect replacements for the originals; virtually all of the content has been accurately transcribed, including the articles missing from the online edition, but the original layout has not been preserved and some of the sidebars are missing. He Qinhua 何勤华, ed. *Faxue de lishi: 1956nian-1957nian* 法学的历史: 1956 年-1957 年 [*The History of* Faxue*: 1956-1957*] (Beijing: Falü chubanshe, 2012); Wang Limin 王立民, ed. *Faxue de lishi: 1958nian* 法学的历史:1958 年 [*The History of* Faxue*: 1958*] (Beijing: Falü chubanshe, 2012).

Sciences Database 中国国家期刊数据库 [NSSD], a government-sponsored, open-access platform.[14] Mindful of that, this author cleared dozens of articles photocopied long ago from those journals out of his files, but not without first confirming the availability of each as a download. Curiously, some of the articles were missing online, and the omissions suggested a pattern. A research project was born.

My awareness that these omissions were worthy of investigation sprang from my grounding in the subject matter. I have read widely in these particular journals and others like them, conducted years of fieldwork in China on the early PRC legal system, including in state archives, and have interviewed dozens of legal cadres who were active in that period, some of whom wrote for *Zhengfa yanjiu* and *Faxue* and participated firsthand in the controversies they documented. The interpretation of the data presented below rests on that foundation.

The first step in the project was to gauge the scale of the omissions by comparing the online holdings against the original paper issues, and recording any discrepancies.[15] Thus, I pieced together a complete set of original paper editions of both journals drawn from the collections of several libraries around the United States, supplemented by targeted acquisitions from China. To ensure commensurability and keep the scope of the project manageable, I decided to confine my analysis to articles published from 1956 through the end of 1958, the only period when both journals overlapped. Fortunately, this also permits one to juxtapose the Hundred Flowers Campaign against the Anti-Rightist backlash that abruptly followed, when the CCP's brief solicitation of earnest feedback switched into searing retribution.

For the years in question, both CNKI and NSSD have uniform holdings, and uniform gaps in coverage. That is to say, they are both missing exactly the same 30 articles from *Zhengfa yanjiu*, and exactly the same 33 articles from *Faxue*, which suggests at least three possible explanations. First, the vendors may have been working from the same incomplete set of source material, or from different sets with coincidentally identical gaps. However, every issue in my sample receives at least some coverage online, which implies that the missing articles were at hand when the issues were originally scanned, and therefore deliberately omitted. Furthermore, the notion that both vendors would have identical gaps purely by happenstance strains credulity. The persistence of that uniformity also defies sound business sense. After all, one would expect a competent vendor to correct inadvertent omissions. Neither one has. To follow up, I used an online customer service form to report several missing articles to NSSD. Within a couple of weeks, a succession of formulaic responses arrived by email: "We're sorry. After many searches, we still cannot find: '(article title).' Thanks for your inquiry!"

The second possibility is that one vendor is simply mirroring the content of the other. At first glance, this appears plausible, but the mirroring would have to be limited in scope because the

[14] The other two vendors either lack coverage (Wanfang Digital Knowledge Service Platform 万方数据知识服务平台), or merely index the articles without offering full-text downloads (Duxiu 独秀).

[15] The comparisons were made in November 2016, and again in February 2017.

two vendors differ in their coverage of other journals from that period. Finally, the simplest explanation consistent with all of the observable evidence is that both vendors are subject to a common affirmative restraint, perhaps an authoritative blacklist or censorship directive imposed from above.

Figure 1: Pages Missing by Issue, Zhengfa yanjiu (1956-58)

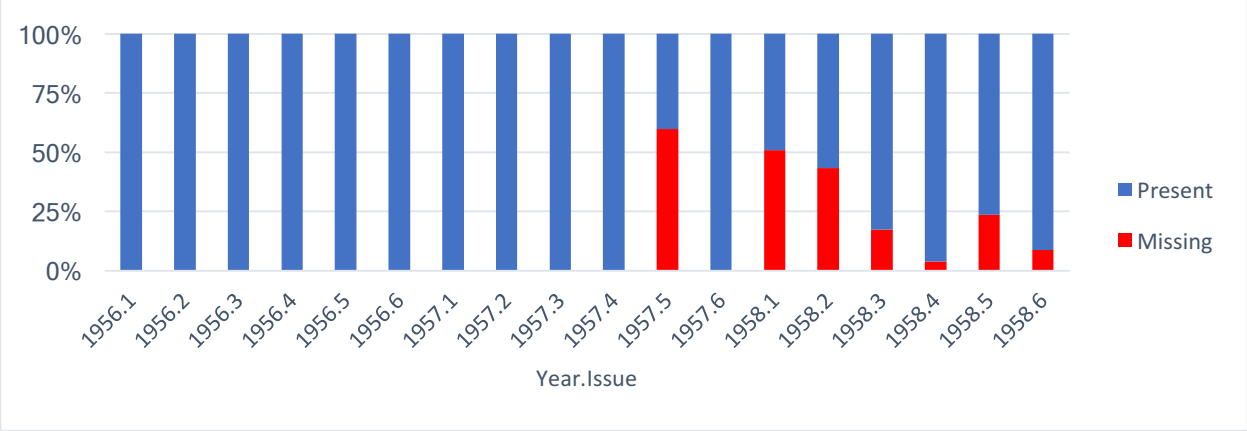
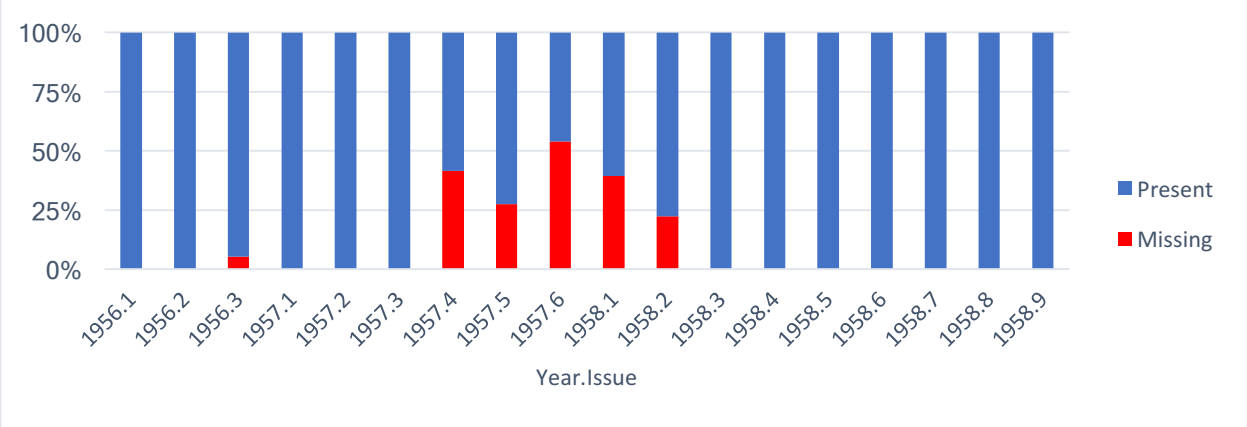


Figure 2: Pages Missing by Issue, Faxue (1956-58)

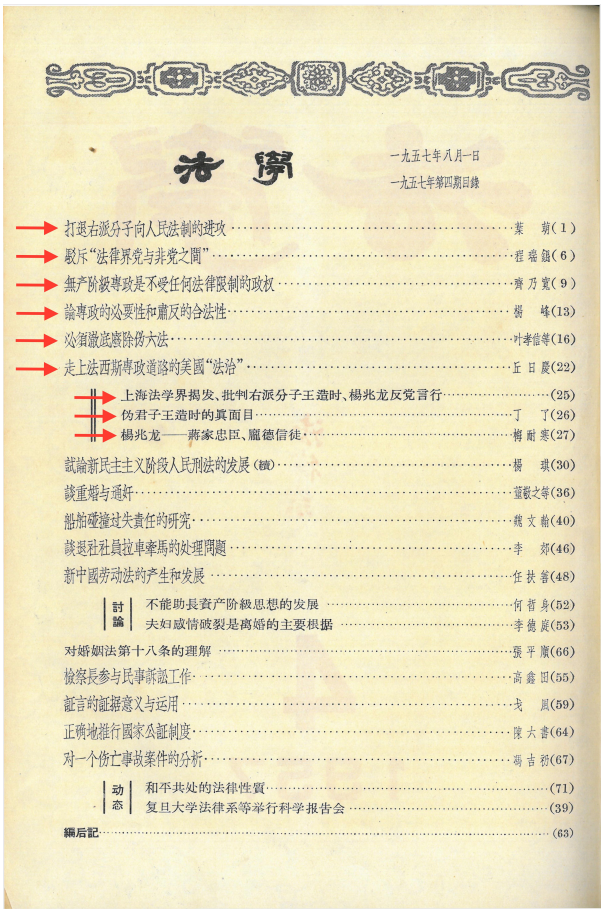


This conclusion is supported by the temporal distribution of the missing articles. Figures 1 and 2 show that the gaps commence abruptly in late 1957 and then taper off as 1958 progresses, corresponding exactly to the first arc of the Anti-Rightist Campaign. The Anti-Rightist Campaign unleashed a ferocious wave of political denunciations and persecutions across China that swept up as many as 700,000 people, and ushered in what Chinese jurists have since called a period of "legal nihilism."[16] The campaign remains a highly sensitive topic in the PRC, and state censorship still restricts inquiries into it.

[16] Zeng Xianyi 曾宪义, "Xin Zhongguo fazhi 50nian lunlüe," 新中国法治 50 年论略 [On Rule by Law in the Past 50 Years Since the Founding of New China] *Zhongguo renmin daxue xuebao* 中国人民大学学报 [*Journal of Renmin University of China*] 6 (1999): 31-41. A handful of scattered articles from earlier issues of these journals have also been omitted online, for example, a 1956 *Faxue* article by the Soviet jurist S.N. Bratus. C.H. Bolatuxi 勃拉图西,

Some of the affected issues are missing around half of their total page count, and the gaps are concentrated almost exclusively at the tops of their tables of contents. In effect, these issues have been decapitated, shorn of their lead articles. Consider the fourth issue of *Faxue* from 1957, which first appeared in August of that year. The table of contents in Figure 3 was scanned from an original paper edition. Red arrows mark the articles that are missing from both online platforms. They account for 30 of the issue's 72 pages. The missing titles appear in translation alongside.

Figure 3: Table of Contents, Original Print Edition of Faxue 1957 (4)



法學

一九五七年八月一日
一九五七年第四期目錄

Articles Missing Online

1. Repulse the Rightist Attack on the People's Legal System
2. Refute "The Legal Profession (is Split) Between Party and Non-Party"
3. The Dictatorship of the Proletariat is a Regime that Does Not Accept Any Legal Limitations

   Sidebar: The Relationship between Yang Zhaolong and Roscoe Pound
4. On the Necessity of Dictatorship and the Legitimacy of the Campaign to Eliminate Counterrevolutionaries
5. We Must Thoroughly Abrogate the Sham (Nationalist) Six Codes

   Sidebar: Excavating Yang Zhaolong's Fascist Roots
6. American "Rule of Law" Travels the Road of Fascist Dictatorship
7. The Shanghai Legal Profession Exposes and Criticizes the Anti-Party Words and Deeds of Rightist Elements Wang Zaoshi and Yang Zhaolong
8. The True Face of Hypocrite Wang Zaoshi
9. Yang Zhaolong: Loyal Minister to the Chiang Family, Disciple of Roscoe Pound

Many researchers interact with online databases primarily through their search engines. Both Chinese platforms offer a conventional suite of search functions to help users locate content, but their search engines are of course blind to the missing content, returning only sanitized

"Guanyu Suweiai yiban faxue lilun kexue gongzuo zhong de jige wenti" 关于苏维埃一般法学理论科学工作中的几个问题 [Several Questions in the Scientific Work on General Legal Theory in the Soviet Union], *Faxue* 法学 [*Law Science*] 3 (1956): 1-4. Analyzing their omission is beyond the scope of this essay.

results and leaving the end user none the wiser. By quietly omitting certain voices and opinions, they are distorting our sense of what the discourse on a given topic actually looked like, where the weight of opinion on it may have been, how that might have changed over time, and why. Not only are they complicit in the intentional misrepresentation of history, but they are also contaminating research based on that source base and violating the trust of their users.

Figure 4: Online Tables of Contents for Faxue 1957(4) (accessed June 7, 2017)

China National Knowledge Infrastructure (CNKI)

序号	篇名	作者	页码
1	讨论新民主主义阶段人民刑法的发展(续上期)	杨琪	30-35+29
2	谈重婚与通奸	董敬之,熊克刚,熊耀武,陈鼎声,刘志舟,许玉霞,夏道平	36-39
3	复旦大学法律系等举行科学报告会		39
4	船舶碰撞过失责任的研究	魏文翰	40-45+51
5	谈退社社员拉车牵马的处理问题	李郊	46-48
6	新中国劳动法的产生和发展	任扶善	48-51
7	不能助长资产阶级思想的发展	何哲身	52-53
8	夫妇感情破裂是离婚的主要根据	李德庭	53-54
9	检察长参与民事诉讼工作	高鑫田	55-59
10	证言的证据意义与运用	戈风	59-63
11	正确地推行国家公证制度	陈六书	64-66
12	对婚姻法第十八条的理解	张平顺	66
13	对一个伤亡事故案件的分析	冯吉礽	67-70
14	和平共处的法律性质——记近年来学术性国际会议对于这一问题的争论	周子亚	71-72

National Social Sciences Database (NSSD)

收录期刊

- 1957年01期
- 1957年02期
- 1957年03期
- 1957年04期
- 1957年05期
- 1957年06期

1957年 第4期 目录导读 整刊阅读

篇名	作者	下载
证言的证据意义与运用	戈风	
正确地推行国家公证制度	陈六书	
对婚姻法第十八条的理解	张平顺	
对一个伤亡事故案件的分析	冯吉礽	
和平共处的法律性质——记近年来学术性国际会议对于这一问题的争论	周子亚	
讨论新民主主义阶段人民刑法的发展(续上期)	杨琪	
谈重婚与通奸	董敬之;熊克刚;熊耀武;陈鼎声;刘志舟;许玉霞;夏道平	
复旦大学法律系等举行科学报告会		
船舶碰撞过失责任的研究	魏文翰	
谈退社社员拉车牵马的处理问题	李郊	
新中国劳动法的产生和发展	任扶善	
不能助长资产阶级思想的发展	何哲身	
夫妇感情破裂是离婚的主要根据	李德庭	
检察长参与民事诉讼工作	高鑫田	

相关期刊

Perhaps this is why they conceal the cull. Users who happen to drill down to the online table of contents for an affected issue will find an unbroken list of articles with no placeholders or notations for missing content, and may consequently never realize that anything is amiss. (Figure 4) On one site, the only hint would be unexplained gaps in the page number sequence of the articles, a detail easy to overlook or miss the significance of. The other site omits even this clue by foregoing page numbers altogether.

Analysis:

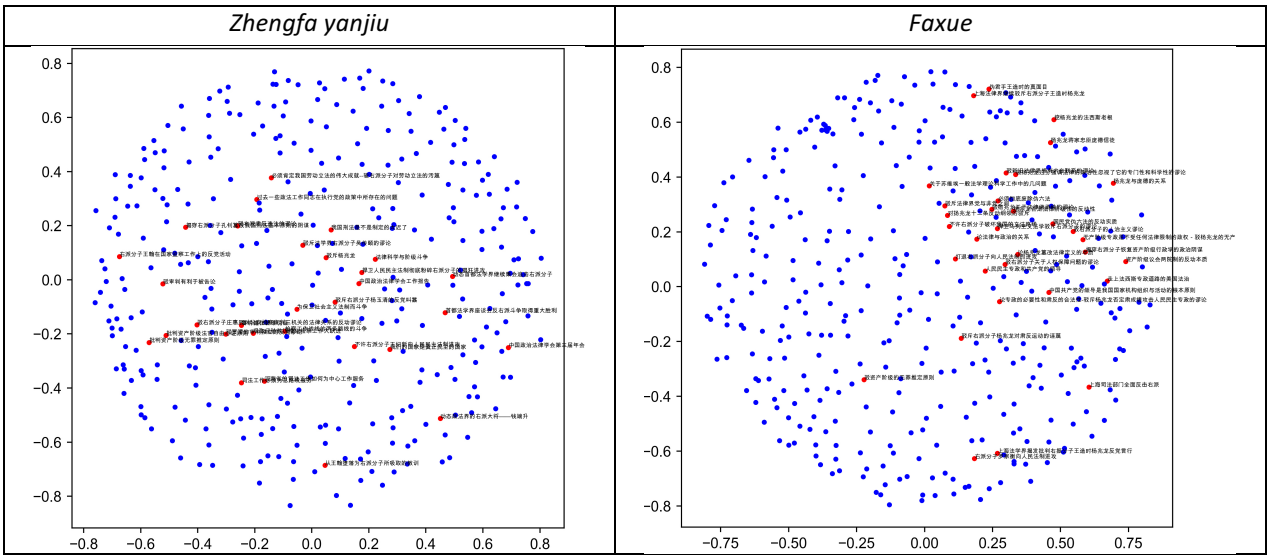
Having pinpointed the anomalies, let us now probe their substance with a mixture of computational and qualitative analysis. In statistical terms, language tends to be noisy and weakly differentiated, particularly when the documents in one's corpora all belong to the same genre, period, and subject matter, such as Chinese law review articles from the late 1950s. The constraints on free expression common to Leninist regimes can magnify this homogeneity. My findings are nevertheless edifying.

My corpora consist of every article and sidebar published in both journals from 1956 through 1958, a dataset comprising 737 documents and nearly four million characters. I began by asking several basic questions about them. Are the online omissions topically stochastic or deterministic? If stochastic, I would expect to see a random distribution of topics across the subsets of missing and included articles, with no obvious features separating one from the other. This would make the case for intentional censorship harder to sustain unless the criteria are essentially arbitrary, or perhaps exogenous to my variables and beyond what they are able to capture. If deterministic, I would expect to see structure in the data, some degree of clustering or lumpiness within the missing and included subsets. Assuming I detected structure, the next step would be to characterize it, to ask which features account for the omission of some articles and not others, and to measure the strength of their effects. Strong, unambiguous findings would not only support the case for deliberate censorship, but also effectively reverse engineer the criteria that trigger it, and hint at the motivations behind it. The scope of my corpora and the way I coded them allowed me to explore these questions across both time and space (Beijing vs. Shanghai).

In Figure 5, the spatial arrangement of the dots expresses the relative proximity of my texts to one another based on their discursive (cosine) similarity.[17] Each blue dot represents a document in the online corpus, and each red dot represents a document missing from it. The red dots also carry labels to identify the names of the specific documents they refer to. One can plainly see that the *Faxue* corpus exhibits sharper internal differentiation than the *Zhengfa yanjiu* corpus, and this result holds across my other statistical tests. Apart from a few stray members, nearly all of the *Faxue* articles missing online inhabit the upper right quadrant of the sphere, which is to say that in spite of variation in their authorship, topics, and wording, as a group they demonstrate an unmistakable coherence. This finding strongly suggests that the omission of texts from the online edition of *Faxue* is not random, but rather involves a discriminating logic, though we cannot yet say what that is.

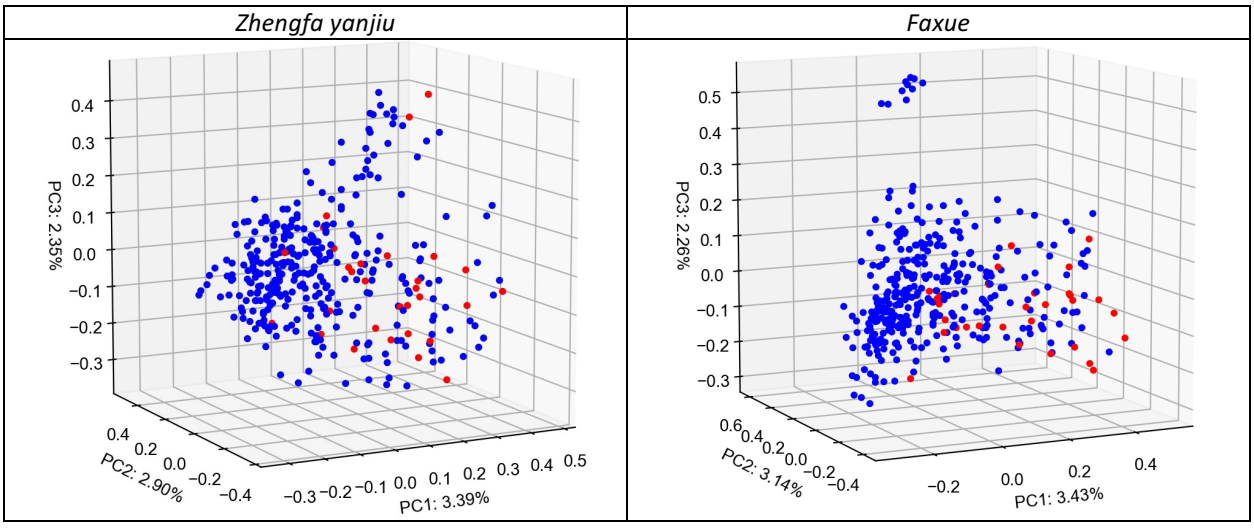
[17] For further information, see the appendix.

Figure 5: Cosine Similarity Measurements



We can uncover hidden differentiation in the *Zhengfa yanjiu* corpus by using a different statistical procedure and adding a third dimension. (Figure 6) Nearly all of the missing documents fall outside of the dense blue core, which indicates that something about them, too, is different, though not to the extent that they cleave cleanly away. Recall, they still belong to a common genre and field of knowledge.

Figure 6: Principal Component Analysis (3D)

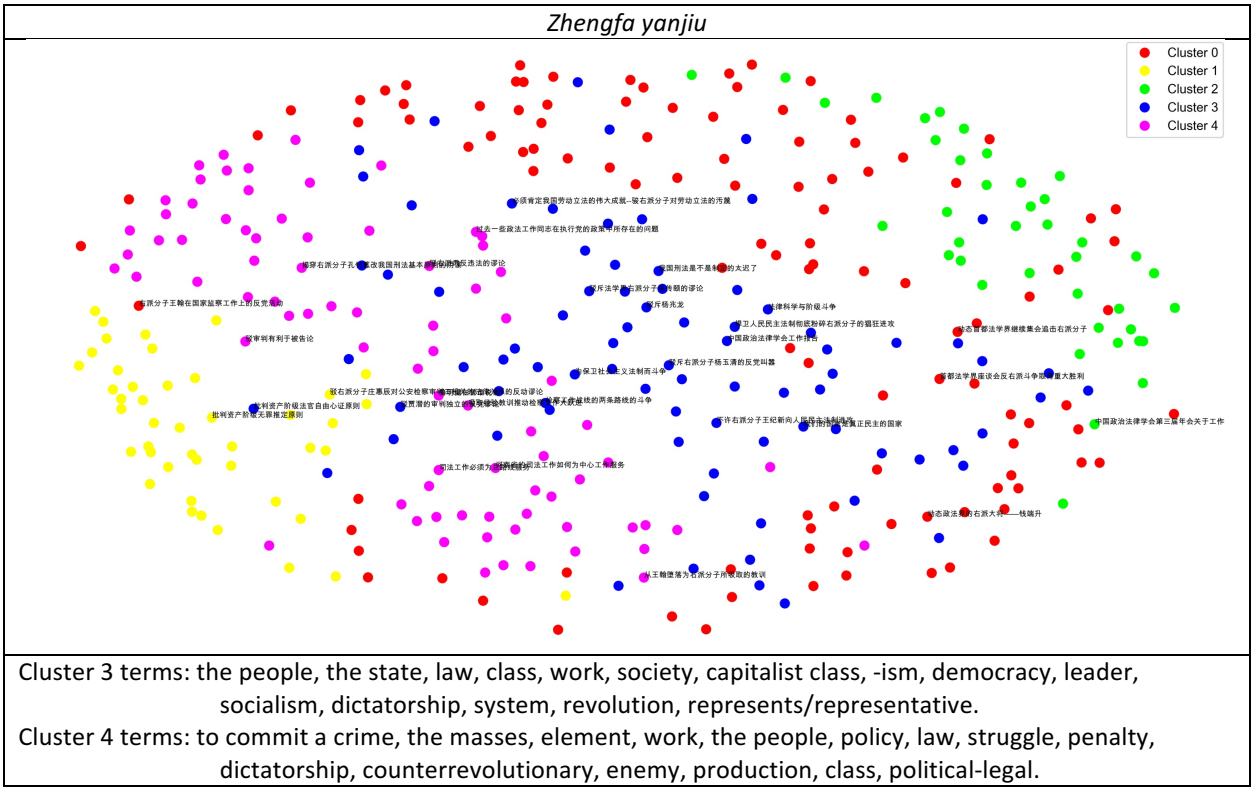


Having established that the omissions are unlikely to be random, we can now try to analyze their characteristics. A procedure called spherical k-means clustering reads a corpus and computes a user-defined number of clusters that the algorithm then populates with documents arranged spatially according to their relative similarity. My implementation also color codes the clusters, and reports back both the terms the algorithm selected (from the corpus) to build

each cluster, and the titles of the articles belonging to those clusters. As with the other calculations, this procedure is blind to the presence or absence of the documents online, and merely labels the omitted documents at the end as a convenience to the user.[18]

Figures 7 and 8 display the results, including the terms that define the clusters where the most omitted documents are found. As before, the dispersion of omitted articles belonging to the *Zhengfa yanjiu* corpus suggests less internal discursive similarity. Fourteen of the thirty missing articles belong to Cluster 3, defined by a basket of terms describing the ideological dimensions of the legal system's relationship to the state and society. An additional nine articles belong to Cluster 4, which centers on criminal law. (Figure 7)

Figure 7: Spherical k-Means Clustering



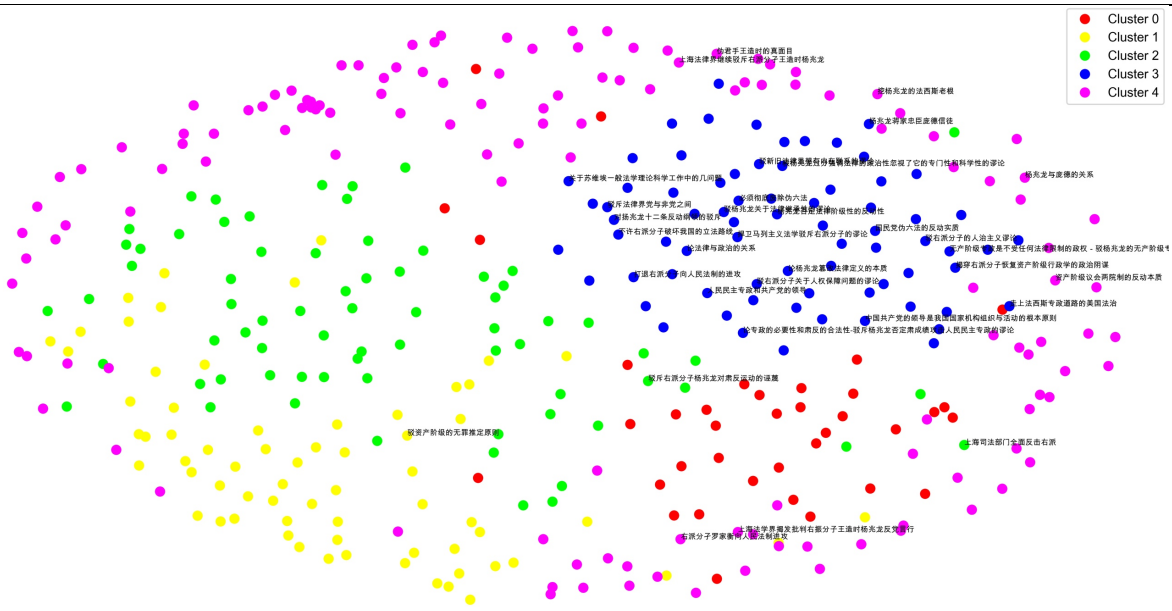
Cluster 3 terms: the people, the state, law, class, work, society, capitalist class, -ism, democracy, leader, socialism, dictatorship, system, revolution, represents/representative.
Cluster 4 terms: to commit a crime, the masses, element, work, the people, policy, law, struggle, penalty, dictatorship, counterrevolutionary, enemy, production, class, political-legal.

For the *Faxue* corpus, the clustering is much tighter, indicating that the omitted documents are internally quite similar to one another. (Figure 8) Moreover, most of the omitted documents fall within a single cluster. The terms defining this cluster overlap significantly with those from the preceding *Zhengfa yanjiu* corpus, with the notable addition of Yang Zhaolong, a prominent Shanghai jurist about whom I will have more to say below. This consistency across corpora begins to dispel the cloud over precisely why some documents have been omitted online, and can guide a trip back to the pertinent texts for a closer, traditional reading.

[18] My implementation of this clustering algorithm adapts code published by Brandon Rose. Rose, Brandon. "Document Clustering With Python." Accessed June 24, 2017. http://brandonrose.org/clustering.

Figure 8: Spherical k-Means Clustering

Faxue



Cluster 3 terms: law, class, the people, the state, dictatorship, society, to rule, democracy, capitalist class, old (*ancien régime*) law, policy, proletarian class, Yang Zhaolong, -ism, socialism.

Let us dive deeper still. A procedure called χ^2 (chi-squared) feature selection measures the strength of the dependence between the appearance of a term in a text and the membership of that text in the omitted class. The higher the score the stronger the correlation. The results are highly informative and allow us, broadly speaking, to reverse engineer the logic behind the omissions.

Figures 9 and 10 list the terms (features) in each corpus most closely correlated with the articles missing online, the relative strength of those correlations, and their respective degrees of statistical significance. Both lists describe similar curves with similar scales. The sole exception relates to Yang Zhaolong 杨兆龙, whose extraordinary score on the *Faxue* list skews its graph. Both curves descend quickly into long tails, which indicates that omission is most tightly associated with just a few features.

Figure 9: χ2 Feature Selection (α=.05)

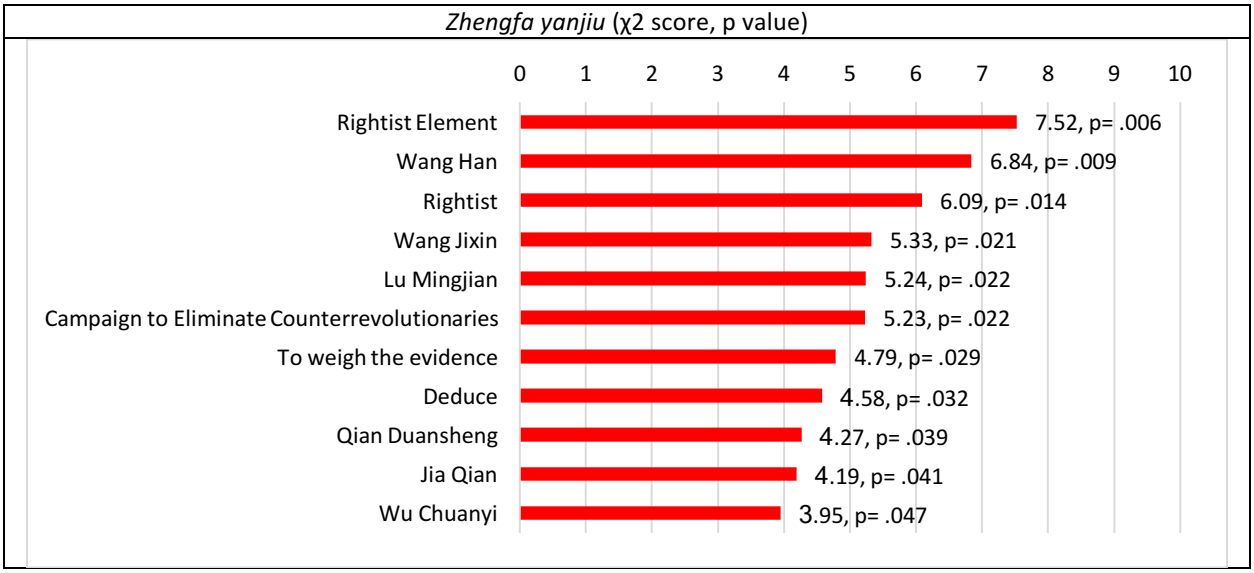
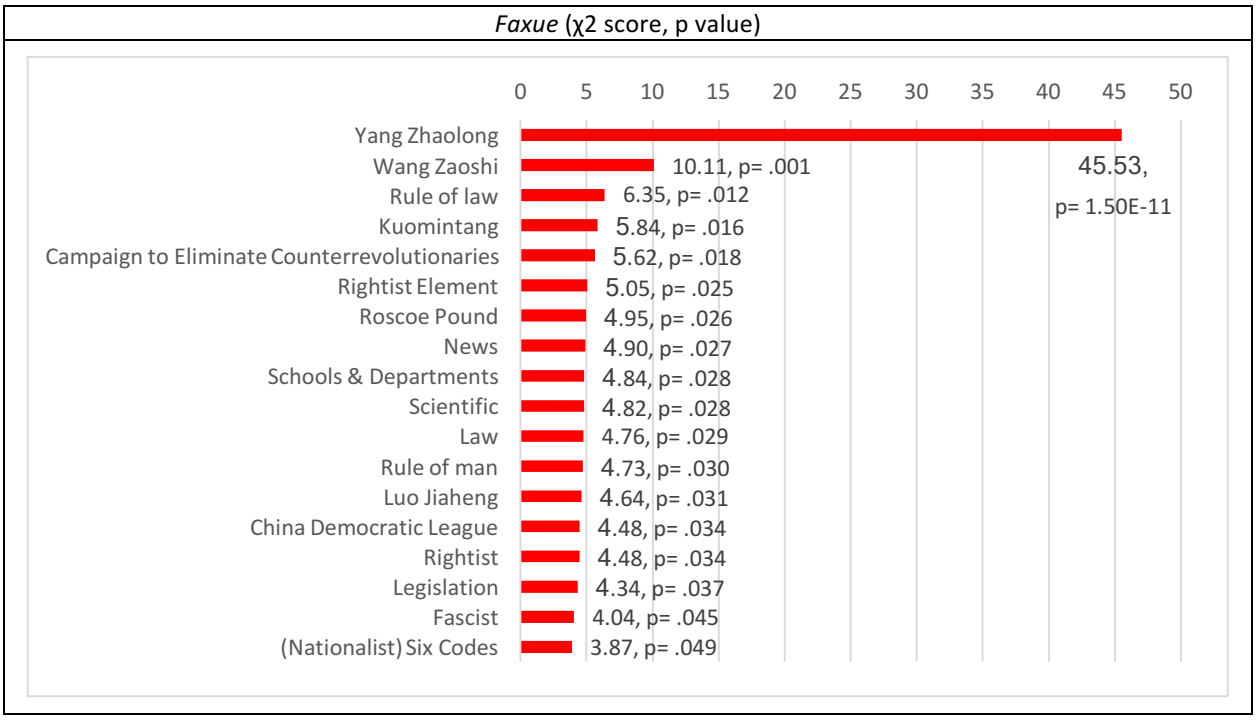


Figure 10: χ2 Feature Selection (α=.05)



On this evidence, I am comfortable attributing the omissions to deliberate, targeted censorship. Note that the selected features differ significantly from one journal to the other, which implies that the censors are responding to different proximate triggers. One can see this plainly with respect to people. Personal names account for more than half of the features on the *Zhengfa yanjiu* list. These include prominent academics with roots in the former Nationalist

establishment, such as Qian Duansheng 钱端升 and Wu Chuanyi 吴传颐, as well as veteran Party cadres and high officials, such as Lu Mingjian 鲁明健 and Jia Qian 贾潜 of the Supreme People's Court. By contrast, the proportion of names on the *Faxue* list is far smaller, and none of the identified individuals were Party members or high officials. The *Faxue* list makes up the difference with features such as "Kuomintang," "legislation" and "rule of law."

Furthermore, the named individuals are geographically segregated, a reflection of how tied each journal was to the concerns of its respective publisher and sponsors. All of the people on the *Zhengfa yanjiu* list were local to Beijing, and all but one on the *Faxue* list were local to Shanghai. The sole exception was Roscoe Pound, the former dean of Harvard Law School, who made the list by virtue of his close past association with the top feature: Yang Zhaolong.

One could surmise from this that the censors are pursuing different aims from one journal to the next but, as the k-means clusters counter-indicate, the contending feature lists actually share a common agenda: burying the evidence each journal respectively betrays of the controversies that rent the emergent PRC legal system. In mid-1957, budding debates over matters such as judicial independence, the transcendence of law over politics and class, the presumption of innocence, and the heritability of law abruptly gave way to vituperative denunciations of those ideas and their sympathizers. Every person listed in Figures 9 and 10 (save Pound, of course) was caught in that shift and persecuted as an example to others. The majority of the remaining selected features also refer to that backlash.

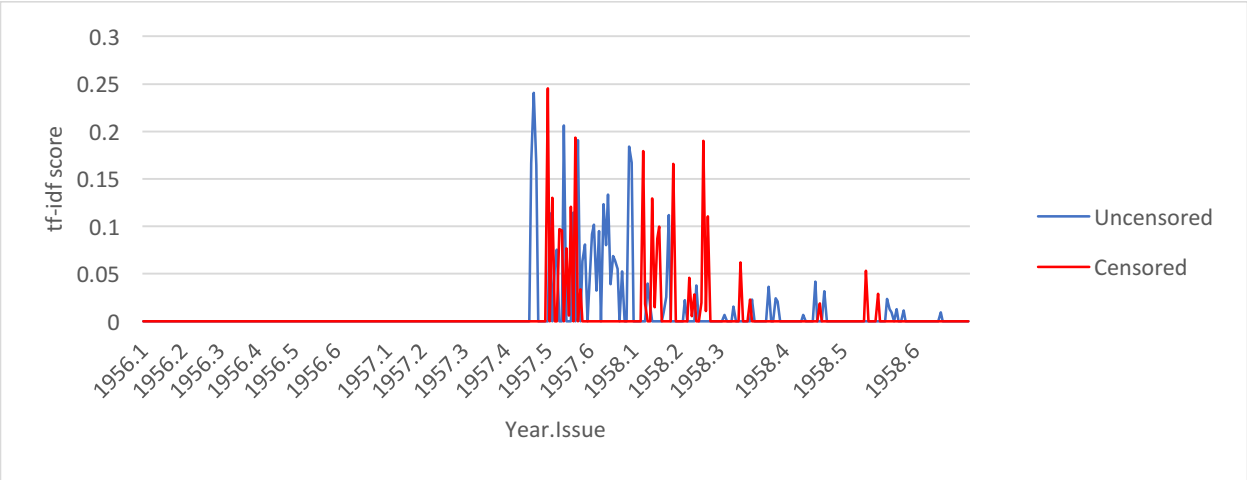
By sanitizing the record, the censors have distorted our understanding of this history, and of the struggles to shape the young PRC legal system. This conveniently absolves the CCP from having to face up to the mortifying contempt for law that defined its early history, and its suppression of those who dissented from that.

To appreciate how the censors are purposefully rewriting history, we can examine the top term by keyness from each journal – "rightist element" 右派分子 (*Zhengfa yanjiu*) and "Yang Zhaolong" 杨兆龙 (*Faxue*) – as well as a term with a lower keyness score that is nonetheless of great significance today: the "rule of law" 法治.[19] Figures 11-14 plot the weight of all three terms over time, arranged in the sequential order of the articles in which they appear, and color-coded according to the censorship status of those articles.[20] This casts a spotlight on exactly which arguments the censors are selecting for and against, and the discursive effect that sorting has.

[19] I sorted the tf-idf matrices of my corpora according to the "keyness" of their 30,000+ unique terms, here defined as the absolute difference between the mean tf-idf scores for a term across all of the articles in the censored class, and all of those in the uncensored class. This provides a ranking of how lopsided the representation of a key term is between the two classes, and therefore helps us to grasp the unique information that is being lost when certain articles are culled. The top 10 terms by keyness for each journal were as follows. *Zhengfa yanjiu*: rightist element, rightist, class, the people, legal system, leadership, struggle, element, capitalist class, judicial. *Faxue*: Yang Zhaolong, Law, class, Wang Zaoshi, dictatorship, capitalist class, the people, Kuomintang, rightist element, democracy.

[20] Weight refers to a term's tf-idf score in a given document.

Figure 11: Rightist Element, Zhengfa yanjiu (1956-58)



Taking Figure 11 as an example, we can see that the term "rightist element" burst into usage quite suddenly at the temporal midpoint of my three-year sample, which coincides exactly with the outbreak of the Anti-Rightist Campaign in mid-1957. After a short period of high intensity, when rightists were hounded, usage dropped off significantly as the campaign ran its course. The graph shows that usage of the term "rightist element" is roughly split between the uncensored and censored classes in this journal, and the term appears with similar weight across both. Unpacking that further is beyond the goals of this essay, but the strength of the technique is that one can easily trace each data point back to its corresponding article for closer inspection. For present purposes, we can simply say that about half of the "rightist elements" in the Anti-Rightist Campaign have gone missing from the online version of this journal. For anyone interested in the history of the PRC, that should set off alarm bells. It calls the reliability and good faith of the online platforms into serious question.

The warping impact of the censors is even more pronounced for the top term in the *Faxue* corpus: Yang Zhaolong (1904-1979). Yang was one of the most outstanding and internationally-respected Chinese jurists of his generation. A graduate of Harvard Law School (SJD '35), he held a variety of high academic and governmental posts in the Nationalist era, including chief of the Ministry of Justice's Criminal Division, where he directed Chinese participation in the Tokyo war crimes trials, and acting Procurator-General. In spite of these positions, Yang, like many committed legal reformers, grew deeply disenchanted with the Nationalist government. In 1949, underground CCP operatives persuaded him to remain on the mainland. In short order, the PRC government appointed him Dean of Soochow University Law School, his undergraduate alma mater, but when it closed the school in 1952, he joined the majority of the former Nationalist-era legal community in professional exile, barred from plying his vocation.

During a brief thaw from 1956 to early 1957, Yang was invited to join Fudan University's law faculty, and the editorial board of *Faxue*. In those few months, he contributed articles to *Faxue* that cogently refuted CCP orthodoxy on the class nature and heritability of law, and on cause

and effect in criminal law.[21] He also drew attention in public fora to the slow pace of codification in the PRC, the low quality of Party legal personnel, and official discrimination against highly-trained, non-Party experts like himself.[22] For his prestige and frankness, Yang paid a heavy price. When the Anti-Rightist Campaign struck Shanghai, a wave of calumny enveloped him.

The original print version of *Faxue* records Yang's remonstrations and the orchestrated campaign against him, but the online edition presents a far different view. The censors have allowed the two conceptual articles Yang wrote for the journal to stand, but not his critiques of the legal system's practical defects or the savage rebuttals he endured. Those have been largely wiped away, taking most of his footprint in this journal along with them. (Figure 12) In fact, the rebuttals against Yang account for the largest cohort of articles censored from *Faxue*, which is why his scores on the χ^2 feature selection and keyness tests eclipse all other terms in my corpora. This is also what endows the censored *Faxue* corpus with such a tight internal discursive coherence. Evidently, the censors wish us to come away with the breathtaking misapprehension that Yang's arguments were mainly theoretical, and well countenanced.[23]

---

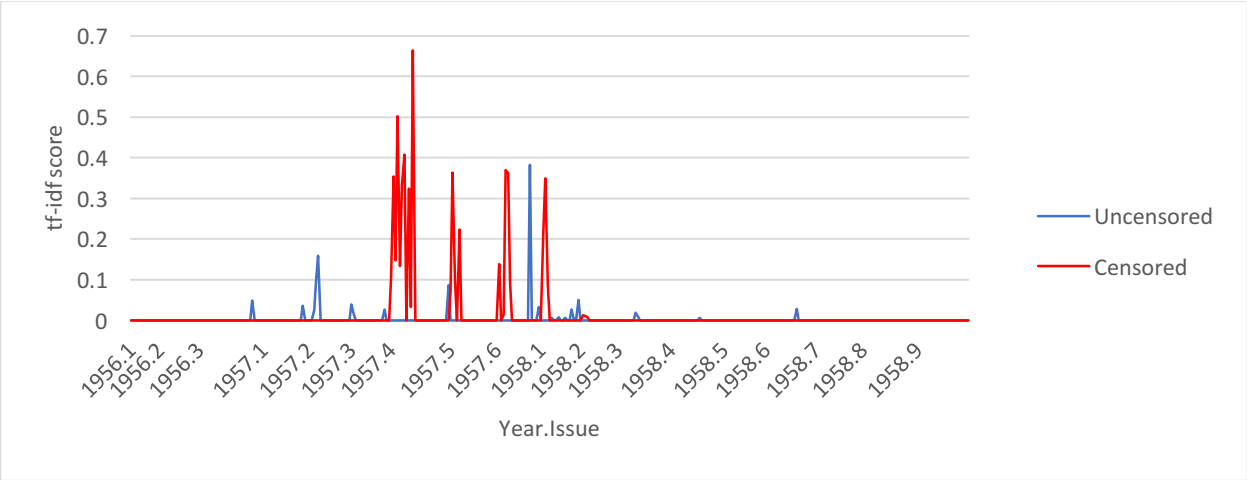
[21] Yang Zhaolong 杨兆龙, "Falü de jiejixing he jichengxing" 法律的阶级性和继承性 [On the Class Nature and Heritability of Law], *Huadong zhengfa xuebao* 华东政法学报 [*East China Journal of Politics and Law*] 3 (1956): 26-34; Yang Zhaolong 杨兆龙, "Xingfa kexue zhong yinguo guanxi de jige wenti" 刑法科学中因果的几个问题 [Several Problems in the Relationship between Cause and Effect in the Science of Criminal Law], *Faxue* 法学 [*Law Science*] 1 (1957): 61-63.

[22] "Peiyang xinsheng liliang, hai you bushao wenti" 培养新生力量, 还有不少问题 [Training Up a New Force, There Are Still Many Problems]. *Xinmin wanbao* 新民晚报 [*New People's Evening Post*], May 4, 1957: 1-2. "Shanghai zhishijie tan guanche baijia zhengming wenti" 上海知识界谈贯彻百家争鸣问题 [Shanghai's Intellectual Community Discusses the Problem of Implementing 'Let One Hundred Flowers Bloom, Let One Hundred Schools Contend']. *Guangming ribao* 光明日报 [*Enlightenment Daily*], May 1, 1957: 2. "Sifa gongzuo 'qiang' gao 'gou' shen, Minmeng shiwei zhaokai sifa zuotanhui pangtingji" 司法工作'墙'高'沟'深, 民盟市委召开座谈会旁听记 [The 'Walls' Are High and the 'Chasms' Deep in Judicial Work, Notes from a Forum Convened by the Municipal Committee of the China Democratic League]. *Xinmin bao* 新民报 [*New People's Evening Post*], May 19, 1957: 1. Yang Zhaolong 杨兆龙, "Falüjie dang yu feidang zhi jian" 法律界党与非党之间 [The Split Between Party and Non-Party in the Legal Community]. *Wenhui bao* 文汇报 [*Wenhui Daily*], May 8, 1957: 2; Yang Zhaolong 杨兆龙, "Woguo zhongyao fadian heyi chichi hai bu banbu?" 我国重要法典何以迟迟还不颁布? [Why After So Long Have China's Key Legal Codes Not Been Promulgated?]. *Xinwen ribao* 新闻日报 [*Daily News*], May 9, 1957: 2-3; Yang, Zhaolong, "Wo tan jidian yijian" 我谈几点意见 [I Discuss Several Points]. *Xinwen ribao* 新闻日报 [*Daily News*], June 6, 1957: 3.

[23] Mei Naihan 梅耐寒. "Guanyu 'fa de jiejixing he jichengxing' de taolun: jieshao Shanghai faxuehui dierci xueshu zuotanhui" 关于'法的阶级性和继承性'的讨论: 介绍上海法学会第二次学术座谈会 [The Discussion on 'The Class Nature and Heritability of Law': Introducing the Shanghai Law Society's Second Academic Forum]. *Faxue* 2 (1957):

Back in the real world, however, they made him the top target of the Anti-Rightist Campaign in Shanghai's legal community, and led to twelve years of imprisonment as a counterrevolutionary.

Figure 12: Yang Zhaolong, Faxue (1956-58)



On this count, Yang Zhaolong is exceptional only in scale, not in kind. The same pattern generally applies across the feature and keyness lists I have computed. Much like the retouched official photo of Mao's funeral, in which gaps in the mourning line belie the erasure of the Gang of Four, the history of this period has been intentionally manipulated to mask elements the authorities would now sooner forget.[24]

Why should censors, and the government that presumably directs them, bother to bury evidence of jurisprudential battles fought 60 years ago? It can only be because those battles bear on issues of cardinal significance today. Since the late 1990s, the CCP has committed itself to constructing the "rule of law" in China, and Article 5 of the state constitution specifically declares the PRC to be a socialist country under the rule of law. But the CCP has an idiosyncratic, exceptionalist understanding of what that means.[25] It has dissociated itself not only from foreign variants of the rule of law, but also from Chinese ones that happen not to be socialist, like those endorsed by Republican governments before 1949, or on Taiwan today.

28–30. Zhang Jinghua 张景华. "Guanyu falü jichengxing zhong de jige wenti" 关于法律继承性中的几个问题 [Several Questions Concerning the Heritability of Law]. *Faxue* 法学 [Law Science] 5 (1957): 18–21.

[24] Marsh, Bill, "Faked Photographs: Look, and Then Look Again." *New York Times*, August 22, 2009: wk4. David King, *The Comrade Vanishes: The Falsification of Photographs and Art in Stalin's Russia* (London: Tate, 2014).

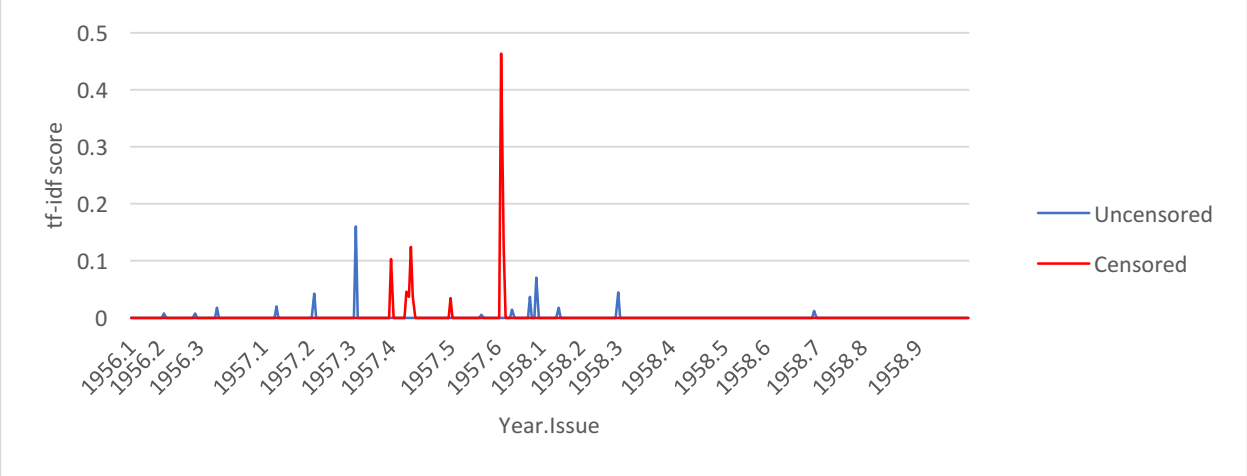
[25] Zhonggong zhongyang zhengfa weiyuanhui 中共中央政法委员会 [Political-Legal Committee of the CCP Central Committee], *Shehui zhuyi fazhi linian duben* 社会主义法治理念读本 [*Reader on the Concept of Socialist Rule of Law*] (Beijing: Zhongguo chang'an chubanshe, 2009).

This leaves it little choice but to look for legitimacy within, and the pickings are slim. From 1957 to 1976, China was rocked by a succession of devastating political campaigns: the Anti-Rightist Campaign, the Great Leap Forward, the Socialist Education Campaign, and the Cultural Revolution. None of these offers a credible pedigree for the current program of "socialist rule of law with Chinese characteristics," and so the early 1950s must carry the burden alone. The arguments raised by Yang and others, and the searing backlash against them subvert this possibility by exposing the gravity of the legal system's congenital defects, and the ferocity of the controversies that seized it. More damning still, they bear witness to where the nation's revered founders actually stood. And in this, they deny the legal system the origin myth it craves.

We can corroborate that conclusion by analyzing how censorship has sculpted the "rule of law." Up until 1953, some senior judicial cadres in the PRC spoke approvingly of the rule of law, but the revolutionary politics of the Judicial Reform Campaign (1952-53) forced them generally to abandon the term and move instead into the safe ideological harbor of Soviet-style "socialist legality." The "rule of law" came to be used derisively, as a marker for the hypocrisy behind which capitalist legal systems cloaked their class domination and exploitation. Furthermore, the values associated with it, such as judicial independence, professionalism, impartiality, and procedural rigor were suppressed as heterodox or counterrevolutionary. The leading academic law journals of the period, *Zhengfa yanjiu* and *Faxue*, document this incontrovertibly, which poses a problem for myth makers today.

Their response should not surprise us. Owing to the ideological climate, invocations of the "rule of law" in these journals are few. That alone should raise a flag. Just the same, censors have manipulated them with care. From my dataset we can reconstruct their original footprint.

Figure 13: Rule of Law, Faxue (1956-58)



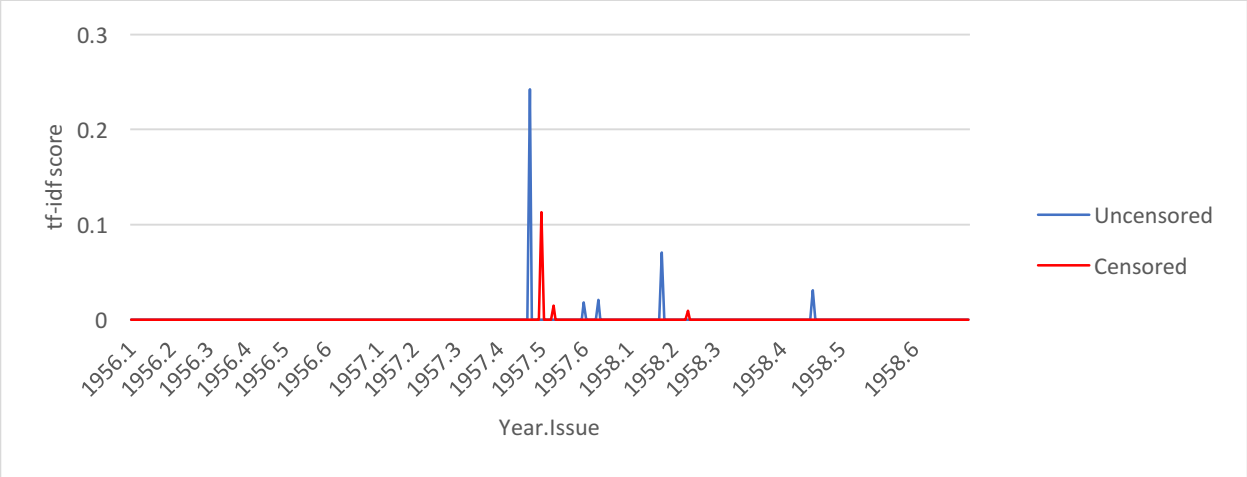
Given the importance PRC authorities now attach to establishing a pedigree for "socialist rule of law with Chinese characteristics," one might expect them to showcase the article in which the

"rule of law" appears with the heaviest weight. It towers above the others. Yet, the opposite is true; it has been censored. (Figure 13) Why?

Perhaps because this article drips with scorn for the concept, and baldy declares that "as far as our socialist nation is concerned, the controversy between the rule of man and the rule of law has no significance whatsoever."[26] "A socialist legal system is not for shackling the limbs of the people. It is mainly directed at all of the enemies of the laboring people. We do not regard law as transcending class, or as deciding everything in perpetuity, but rather grasp it as one of the principal means for effecting the dictatorship of the proletariat."[27] Clearly, this is not the socialist rule of law tradition today's authorities are looking for.

What about the peak *Faxue* article in the uncensored class? This article considers the doctrine of the "rule of law" proposed by the third century BCE philosopher Han Fei.[28] His is a stern, top-down vision of law closer in the Western canon to classical *rechtsstaat*, or "rule by law." It emphasizes the interests of the state over the individual, and grants the sovereign unquestioned supremacy over subjects. Its synergy with modern authoritarianism is evident, and as a source of legal tradition with indigenous roots deep enough to resist the intrusion of modern Western competitors, the CCP could hardly do better. Hence, it not only avoids the scalpel, but dominates the (blue) sanitized discourse on the "rule of law" remaining online.

Figure 14: Rule of Law, Zhengfa yanjiu (1956-58)



By contrast, in *Zhengfa yanjiu*, the peak article remains uncensored. (Figure 14) An account of a four-day struggle meeting organized in Shanghai at the start of the Anti-Rightist Campaign, this

[26] Zhai Tingjin 翟廷瑨, "Bo youpai fenzi de renzhi zhuyi miulun" 驳右派分子的人治主义谬论 [Refuting the Absurd Rightist Element Theories on the Doctrine of the Rule of Man], *Faxue* 法学 [*Law Science*] 6 (1957): 8.

[27] Zhai, "Bo youpai fenzi de renzhi zhuyi miulun," 11, 43.

[28] Mou Jie 缪杰, "Lüelun Han Fei de fazhi zhuyi" 略论韩非的法治主义 [Introduction to Han Fei's Doctrine of the Rule of Law], *Faxue* 法学 [*Law Science*] 3 (1957): 9-15.

article claimed that "at the meeting many people forcefully refuted Wang Zaoshi's slander against the legal system, took on his absurd notion of [China] 'stressing rule of man while slighting the rule of law', and listed numerous hard facts to explain that China has the best rule of law. Luo Zhufeng said: 'the quality of rule of law can be measured on the basis of social order and democratic rights, and in China the stability of the social order and the realization of democratic rights has historically never been greater.'"[29] *This* is the native socialist rule of law tradition the censors wish us to discover, and they have thoughtfully cleared a path to it by winnowing away all rivals.

Thus far, I have spoken only of two law journals from a narrow slice in time, and the attentive reader might wonder if the case I am making applies more broadly. It does. Let us look first at another law journal from the same period. *Renmin sifa gongzuo* 人民司法工作 [*People's Judicial Work*] debuted in 1957 under the dual sponsorship of the PRC Ministry of Justice and the Supreme People's Court. The following year, its name changed to *Renmin sifa* 人民司法 [*People's Judicature*], and the Supreme People's Court took sole custody. The journal had a restricted (*neibu*) circulation, and the battles fought in its pages, which mirrored those in *Zhengfa yanjiu* and *Faxue*, were therefore hidden from public view. This was where the judiciary spoke privately to itself.

To the best of my knowledge, no library in the United States has physical copies of the 1957 and 1958 volumes of this journal, which makes the PRC online edition precious. Unfortunately, only one of the four online vendors offers full-text articles from any of these volumes, and its coverage is fragmentary. All of 1957 is missing, comprising 12 issues, or 292 articles and 513 pages of text. The online edition also omits 135 articles from 1958, including all of issues 1-3 and 5-10, or 40 percent of the articles for the year. This virtually zeroes out the record of the journal's participation in the Hundred Flowers Campaign and the outbreak of the Anti-Rightist Campaign.

It is not hard to understand why. Many of the missing articles pillory high-ranking rightists in the judiciary for their "anti-Party" attachments to principles like judicial independence, the burden of proof, and weighing the arguments from both sides in a case. That history now embarrasses the CCP. Consider, for example, this excerpt from a page one editorial in one of the censored issues, again the official organ of the Supreme People's Court:

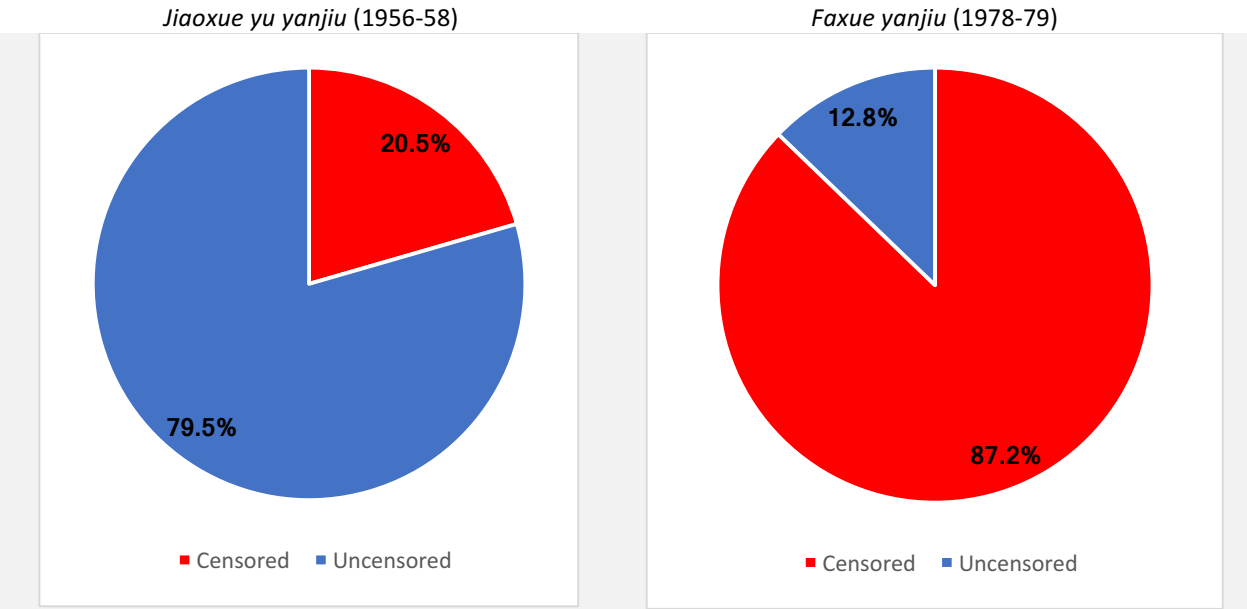
> Rightists wanted to use "adjudicatory independence" to oppose the leadership of the Party over judicial work…Is this rule of law or rule of man? We have rule of man. What

[29] "Shanghai faxuejie zhankai fan youpai fenzi de douzheng" 上海法学界展开反右派分子的斗争 [Shanghai's Legal Community Launches the Struggle Against Rightist Elements], *Zhengfa yanjiu* 政法研究 [*Political-Legal Research*] 4 (1957): 53.

is called the rule of man is relying on Party leadership, relying on the masses, giving free rein to the dynamism of human subjectivity.[30]

Fair enough, but one might still wonder if the censorship is limited to law journals. To be certain, I checked issues of a leading interdisciplinary social science journal: *Jiaoxue yu yanjiu* 教学与研究 [*Teaching and Research*]. The same pair of online platforms (CNKI and NSSD) offer full-text coverage, and their holdings and omissions are again identical. Nearly 20 percent of the total page count published from 1956 through 1958 is censored online, including five complete issues and approximately two dozen additional scattered articles on topics ranging from philosophy to the arts, and village life. (Figure 15) Additionally, the publisher, People's University, offers the journal on its own servers, and omits the same issues, which would be consistent with a uniform censorship directive from above.[31]

Figure 15: Censorship Across Other Journals, by Page Count



A skeptic might further wonder if the censorship I have found simply reflects the Party's special sensitivities about the Anti-Rightist period. Not at all. The Anti-Rightist Campaign is merely one example of a much more extensive problem; the CCP's insecurity and deep contempt for history guarantee that no topic or period is safe from similar manipulation. We have only to turn to the reconstruction of the legal system during the late 1970s for proof of that.

The first academic law journal to appear after the Cultural Revolution was *Faxue yanjiu* 法学研究 [*Studies in Law*], the successor to *Zhengfa yanjiu*. Seven issues were published from late

[30] "Ba xianchang huiyi de jingshen dai dao shiji gongzuo zhong qu" 把现场会议的精神带到实际工作中去 [Carry the Spirit of the This Meeting to Practical Work], *Renmin sifa* 人民司法 [*People's Judicature*] 19 (1958): 4.

[31] The journal website is: http://xsqks.ruc.edu.cn/Jweb_jxyyj/CN/article/showOldVolumn.do

1978 to the end of 1979, the first two of which were trial balloons with restricted, internal circulation. Publication of the journal continues to this day under the sponsorship of the Institute of Law of the Chinese Academy of Social Sciences in Beijing.

At the time, Deng Xiaoping was still consolidating power. To check his rivals on the left and to lay the groundwork for the trial of the Gang of Four, he permitted legal discourse a degree of latitude not seen since the mid-1950s. This coincided with the Democracy Wall movement, which took popular political protest and dissent to the streets. The 1978 and 1979 issues of *Faxue yanjiu* document that brief opening.

Many of the articles in those issues reflect on the sorry state of the legal system, which they blame on the Gang of Four. But some hint at wider culpability and express views that have since fallen out of favor. For instance, one article frankly observes: "it is obvious to everyone that over the years the doctrine of the rule of man harmed Chinese political life. Legal nihilism and the doctrine of the rule of man gained ground as early as the late 1950s. Through the early 1960s, the standpoint of favoring the rule of man and rejecting the rule of law was extensively propagated."[32] Another article asserts that Lin Biao and the Gang of Four "not only inherited the dregs of historical Confucian and Legalist thought, but also promoted them by usurping power. For a time, revolutionary leaders were deified and feudal-fascism ran wild, concealed under the heavy curtain of the rule of man."[33]

Such opinions may have been convenient for the proximate aim of discrediting the Gang of Four, but they are heterodox today, and censorship provides a tempting way out of that quandary. Censorship frees the Party to argue instead that the roots of socialist rule of law in the PRC run deep, and that Confucian and Legalist philosophy are actually excellent sources of legal tradition for it.[34] If they have done their jobs well, nothing in the online record will disprove this, or indicate that orthodoxy once dictated otherwise.

A check of the record reveals that only one of the four major online vendors in the PRC offers any full-text coverage of *Faxue yanjiu*'s 1978 and 1979 issues, and the holdings are thin. More than 87 percent of the total page count is missing, including the pieces cited above, and all six issues from 1979. For the subscriber, it is almost as if the discourse that jumpstarted academic legal scholarship in the post-Mao PRC, and captured the promise of that pregnant moment, never existed. (Figure 15)

Admittedly, no repository of the past can achieve completeness, but something insidious lurks behind the insufficiencies of these particular platforms. They are exploiting the malleability of

[32] Gu Chunde 谷春德, Lü Shilun 吕世伦, and Liu Xin 刘新, "Lun renzhi yu fazhi" 论人治与法治 [On the Rule of Man and the Rule of Law], *Faxue yanjiu* 法学研究 [*Studies in Law*] 5 (1979): 27.

[33] Zhang Jinfan 张晋藩, and Zeng Xianyi 曾宪义, "Renzhi yu fazhi de lishi pouxi" 人治与法治的历史剖析 [Historical Analysis of the Rule of Man and the Rule of Law], *Faxue yanjiu* 法学研究 [*Studies in Law*] 5 (1979): 37.

[34] Zhonggong zhongyang zhengfa weiyuanhui, *Shehui zhuyi fazhi linian duben*.

the digital record to turn history on its head, portraying current legal reform as resting on a venerable founding vision.[35] This is originalism by censorship, and while irksome enough for foreigners, its primary audience is surely Chinese. It aims to control China's future by shaping consciousness of its past. With the passage of time, detecting, exposing and debunking such ploys gets ever harder and more expensive, and this alone may be enough for them to succeed.

The hazards described here undoubtedly apply everywhere the digital turn takes hold. If knowledge is power, then the consolidation of information in the hands of a comparatively small number of digital providers, and the conversion of our libraries from redundant, fault-tolerant repositories of tangible objects into passive links in a centralized distribution chain controlled by those providers is worrying. Increasingly, we are dependent on the good faith of their stewardship, and vulnerable to the political, regulatory, commercial and licensing terms that may impinge upon it. Moreover, we are voluntarily surrendering the evidence necessary to independently monitor their performance and hold them to account.

I cannot say if the censorship this study investigates is algorithmically-driven, but to explore that possibility I used a multinomial naïve Bayes classifier that employs machine learning to build a censorship model, which I then tested against a known control set. This resembles the technology used to identify and filter spam email. The model regularly achieved more than 70% accuracy, a respectable level given the small size of the censored class, and the common genre to which all the articles belonged. These preliminary results, performed on a small dataset with a modest laptop computer, show that from a distance I can train a machine to approximate the choices the censors are actually making about these journals, which further underscores the likelihood that there is a deterministic, discoverable logic guiding their decisions.

Considering the scale of their responsibilities, the censors may already be using similar technology. With the resources available to a state or to a well-financed private firm, the possibilities are chilling. For instance, simply by fine-tuning any of the 30,000+ parameters in my dataset, an enterprising censor (or hacker) could fabricate bespoke corpora with surgical precision as often as they like, without leaving behind any fingerprints, and subscribing libraries would in turn faithfully disseminate each successive edit. Likewise, Chinese state archives, many of which are aggressively digitizing their collections, could use such methods to modulate the accessible record. Researchers would then integrate these dynamically curated source bases into their scholarship and propagate their biases around the world, much as in a classic disinformation campaign, only executed with negligible transaction costs and seamless efficiency.

[35] For more on this topic, see Glenn Tiffert, "Socialist Rule of Law With Chinese Characteristics: A New Genealogy," in *Socialist Law in Socialist East Asia*, ed. Hualing Fu, John Gillespie, Pip Nicholson, and William Partlett (New York: Cambridge University Press, forthcoming).

Solutions:

One obvious response would be to bring the missing texts to the attention of the vendors, and request remediation. My efforts as an individual consumer met with disappointment. But a subscription to one of these Chinese digital platforms can cost a research university tens of thousands of dollars, and the possibility of collective action by a group of them to demand a better product could prod the vendors in the right direction. At the same time, the evidence also suggests that the decision may not be theirs to make, and persuading the Chinese government to reverse itself is a much taller order.

Another response would be to republish the censored texts, but there is a vexing obstacle to that. Under current U.S. law, all of the journals examined in this study are under copyright and will remain so until 95 years after their date of original publication, i.e. at least until the early 2050s.[36] This forecloses many of the options one can take to circumvent their censorship since any unauthorized republication of the sources would constitute infringement. As with conventionally orphaned works, few institutions would risk liability, and the Chinese rights holders could use U.S. law as a club to enjoin any that did.

Thus, intellectual property law perversely abets the manipulation of history by an authoritarian regime. That regime need only wait for the few surviving physical copies of the censored texts to crumble or fade into obscurity. Many libraries are actually accelerating those outcomes by deprecating legacy volumes on their own initiative, oblivious to the stakes. In fact, all copyrighted works are susceptible to this morbid waiting game. For example, the iconic film *River Elegy* (河殇 *He Shang*), banned in the PRC since 1989, languishes in much the same position. A concerted effort outside of China to digitize these materials for the sake of preservation, and to sequester them until they fall out of copyright decades from now is perhaps the most straightforward solution, though a deeply unsatisfying one.

Technology has other parts to play. In this study, I used computational tools to demonstrate that the dystopian possibilities of the digital turn are already among us. There is a certain irony in that. These tools equipped me to analyze the relationships among documents at a scale and resolution beyond my own cognitive abilities. But rather than use them as replacements for traditional methods, I employed them complementarily -- as instruments for hypothesis testing, to pull interesting features out of my corpora for closer interpretive scrutiny, and to check my subjective judgments empirically. The tools were also invaluable for analyzing the alternate histories created by the censors, assessing what has been lost, distorted, and fabricated by their work, reverse engineering its logic, and reconstituting the record they manipulated. In short, technology elevated my conventional skills as a historian.

[36] 17 U.S.C. §104A; Elizabeth Townsend Gard, "Vera Brittain, Section 104(a) and Section 104A: A Case Study in Sorting Out Duration of Foreign Works Under the 1976 Copyright Act," *7th Annual Intellectual Property Scholars Conference, DePaul University College of Law* (2007): n.p.; Tyler T. Ochoa, "Copyright Protection for Works of Foreign Origin," in *The Internationalization of Law and Legal Education*, ed. Jan Klabbers, and Mortimer Sellers (London: Springer, 2008): 167-190.

It can also prevent the cancer this study has exposed from spreading. Already, the digital turn is sparking cynicism in society about the authenticity of the knowledge we consume and produce. A mere whiff closer to home of the practices this study has documented could enflame that crisis of credibility and further erode the trust necessary for robust scholarship and democratic practice.[37] It is therefore incumbent on knowledge creators, rights holders, digital providers and the institutions that subscribe to their products to design and consistently implement a set of best practices to uphold the integrity of their holdings, transparently log omissions and modifications, and defend against tampering at the levels of the individual character, document and corpus. We can no longer passively trust, we must also verify. A variety of solutions, such as digital signatures or blockchain certification, with logos signifying validated standards compliance, are potentially available to meet those needs.[38] The danger is real. Never before has knowledge been prone to such sweeping and supple manipulation. History hangs in the balance.

[37] Samuel C. Woolley, and Philip N. Howard, "Computational Propaganda Worldwide: Executive Summary." *Computational Propaganda Research Project* Working Paper No. 2017.11 (2017): accessed June 21, 2017, http://comprop.oii.ox.ac.uk/2017/06/19/computational-propaganda-worldwide-executive-summary/. Environmental Data and Governance Initiative. "The EPA Under Siege." (2017): Accessed June 30, 2017. https://100days.envirodatagov.org/epa-under-siege.html.

[38] Portico (http://www.portico.org), a non-profit, digital preservation service that partners with libraries and publishers to preserve the scholarly record, offers a promising partial solution for committed digital content no longer available from the publisher or any other source, but its standard licensing agreements do not safeguard against the full range of practices described in this study, particularly censorship at the level of individual articles.

Appendix: Tools and Methods

Preparing the documents for the analysis undertaken in this study required several, labor-intensive steps, the most basic of which was converting images into text with a high degree of accuracy.[39] To maximize fidelity, I used only original print editions, not reprints or reproductions. Every page of every issue in the three-year sample, more than 2,000 in total, was scanned at 600 dpi, in greyscale without any lossy compression, resulting in nearly 20 gigabytes of data. Second, I used a commercial optical character recognition (OCR) package to convert those scans into plain text, and I then sliced the output into individual files, one for each article. My final corpora consisted of 356 files from *Zhengfa yanjiu* and 381 files from *Faxue*, or nearly four million characters in total.[40] Both vendors are currently censoring approximately eleven percent of the total page count in my three-year sample, though that share exceeds 50 percent for certain issues straddling the 1957/58 divide.

Third, university-educated, native speakers of Chinese compared my *Faxue* corpus against the original documents, character by character, to establish the reliability of the OCR. The test set comprised 127 pages (approximately 207,000 characters) from the original issues, and averaged a remarkable 99 percent accuracy, easily sufficient for my purposes. Time and funding precluded an equally exhaustive test of my *Zhengfa yanjiu* corpus, but spot checks suggested similar accuracy.

Fourth, I wrote several programs in Python, the first of which stripped my corpora of semantically expendable characters, such as punctuation, alphanumerics, Cyrillic, gremlins, and whitespace. This reduced each file to an unbroken stream of Chinese characters, which was then fed to a segmentation algorithm that tokenized the texts. Unlike Western languages, Chinese does not separate words with whitespace. The segmentation algorithm performs this step, which is required by the natural language processing (NLP) routines that comprise conventional computational text analysis.[41] It bears mentioning that the texts in my corpora were originally published just as the PRC was moving from traditional Chinese characters to the

[39] My workflow relied on the following software: MacOS (10.12.5), ABBYY FineReader (12.1.6), BBEdit (11.6.5), Anaconda3 (3.4.1), Scikit-learn (0.18.1), Java 8 (131), Stanford Chinese Segmenter (3.7) with the ctb standard, Mallet (2.08), and Microsoft Excel (15.35).

[40] One could arrive at a slightly different count, depending on how one divides sidebars and fora with contributions from multiple authors, but the key point is that every character was captured.

[41] My methodology and workflow were informed by various texts in computational linguistics, natural language processing, and digital history, including: Charu C. Aggarwal, and Chengxiang Zhai, eds. *Mining Text Data* (New York: Springer, 2012); Steven Bird, Ewan Klein, and Edward Loper, *Natural Language Processing With Python* (Cambridge: O'Reilly, 2009); Michael C. Hout, Megan H. Papesh, and Stephen D. Goldinger, "Multidimensional Scaling," *Wiley Interdisciplinary Review of Cognitive Science* 4, no. 1 (2013): 93-103; David Mimno, "Computational Historiography: Data Mining in a Century of Classics Journals," *ACM Journal of Computing in Cultural Heritage* 5, no. 1 (2012): 3:1-3:19; Jason D.M. Rennie*, et al.*, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," *Proceedings of the Twentieth International Conference on Machine Leaning* (2003): 616-623; David Underhill*, et al.*, "Enhancing Text Analysis Via Dimensionality Reduction," *IEEE International Conference on Information Reuse and Integration* (2007): 348-353. Portions of my code adapted examples shared by Paul Vierthaler, Brandon Rose, and countless contributors to Stack Overflow.

simplified set used today, and the articles consequently contain a transitional mixture of both. This confused the segmentation algorithm, and necessitated the preliminary measure of converting both corpora uniformly to simplified characters.

Fifth, the segmentation algorithm performed poorly for named entities, particularly personal names, which can be idiosyncratic. For example, surnames were commonly dissociated from given names. I therefore manually compiled a dictionary of several hundred entries comprising every author, as well as prominent organizations and individuals mentioned in my texts. A Python program then scanned the texts for occurrences of these names and reconstituted them correctly.

Sixth, I compiled a series of metadata files necessary for the analytics I intended to perform. These files are essentially spreadsheets (.csv) with columns for the filenames, article titles, author names, index number (year-issue#-article#), and censorship status (Y or N) of every article in various slices of my corpora. I built metadata files for the *Zhengfa yanjiu* corpus, the *Faxue* corpus, a combined corpus, and yet another combined corpus that substituted the city of publication (B[eijing] or S[hanghai]) for the censorship field. Finally, after approximately five months of preparation, the corpora were ready for analysis.

To perform the analytics, I wrote another Python program that stripped out stop words from each pre-processed corpus and then transformed them into a series of tf-idf matrices against which I could run a battery of statistical tests.[42] Briefly, these matrices describe each of the 737 documents in my corpora in more than 30,000 dimensions, where each dimension corresponds to a unique term in a document (e.g. "judge," "Yang Zhaolong," "rule of law," etc.). For the graphs, I then applied a variety of standard techniques to reduce that space to two or three abstracted dimensions that nonetheless retain sufficient information about the documents to represent their essential relationships in a manner that humans can visually comprehend.[43]

[42] By definition, stop words do not carry information significant to a given query, usually because of their high frequency. The archetypical example in English would be "the." Filtering them out removes chaff from the data and reduces processing time, without materially affecting results. A term frequency-inverse document frequency (tf-idf) matrix is a standard statistical representation of a corpus, in which each term in every document is assigned a weighted value based on its frequency in that document and its distribution in the corpus at large. The basic idea is that a term that appears often in a small fraction of the corpus is more informative and of greater selection value than a term that is distributed evenly. The normalized tf-idf values were computed using the *TfidfVectorizer* class in Scikit-learn 0.18.1, according to the following function: $tfidf_{(t,d)} = tf_{(t,d)} \times (log \frac{1+n_d}{1+df_{(d,t)}} + 1)$.

[43] These techniques included t-SNE analysis (TruncatedSVD), principal component analysis, Euclidean distance calculations (MDS), scree plots, cosine similarity calculations (MDS), spherical k-means clustering (MDS), hierarchical clustering analysis (Ward dendrogram), and naïve Bayes classification (multinomial). I also performed χ^2 feature selection, and topic modeling (Mallet, LDA) on my corpora.