Science of Law Journal (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/law.2025.040604
ISSN 2616-2296 Vol. 4 Num. 6

A Comparative Study of Legal Risks and Compliance Frameworks for Web Crawling Technologies in China and Singapore: An Empirical and Doctrinal Analysis in the Age of AI

Ji Jiaqi

Nanyang Technological University, Singapore
13810401561@163.com

Keywords: Web Crawling, Unfair Competition, China Judgments Online, Empirical Legal Study, Data Rights, Compliance Frameworks

Abstract: Automated web crawling plays an important role in search engines, digital platforms, financial analysis, and the development of large AI datasets, but it also raises increasing legal concerns as data becomes more valuable. This study offers the first empirical and comparative analysis of web-crawling regulation in China and Singapore by examining over 500 Chinese court decisions involving issues such as unauthorized access, large-scale data scraping, unfair competition, copyright disputes, and the bypassing of technical protection measures. Using methods including Pareto charts, correlation analysis, and network modeling, the study finds that Chinese courts usually take a flexible and market-oriented approach, balancing platform interests with the legitimacy of accessing public data. In contrast, Singapore applies a clearer and more access-focused legal framework based on the Computer Misuse Act, the Personal Data Protection Act, and a limited text-and-data-mining exception under copyright law, which sets strict boundaries while offering only narrow support for research. The comparison shows that China's system is adaptable but fragmented, while Singapore's framework is consistent but relatively rigid, and the study suggests that combining clarity in access rules with competition-based reasoning could help improve cross-border regulation of automated data collection in Asia.

1. Introduction

1.1 Background: Web Crawling, Data Governance, and Divergent Regulatory Models in Asia

Automated data collection, widely referred to as web crawling or web scraping, has become a foundational mechanism for data-driven innovation across contemporary digital economies. It supports the functioning of search engines, recommendation platforms, financial risk modeling systems, online marketplaces, and the construction of large-scale datasets for machine learning [1]. As digital infrastructures expand and the significance of data intensifies, automated collection technologies determine how information circulates, how digital markets operate, and how artificial intelligence systems acquire the materials needed for training and evaluation. In China's rapidly

evolving digital environment, the use of crawling technologies has grown sharply across the commercial, research, and regulatory sectors, making legal clarity in this field increasingly essential [2].

China, however, lacks a unified legal statute governing access to online data. Instead, the Anti-Unfair Competition Law, Copyright Law, Cybersecurity Law, Personal Information Protection Law, Criminal Law, and Civil Code jointly regulate fragmented aspects of crawling-related conduct. Courts are therefore required to interpret overlapping legal regimes while balancing platform control, user interests, technological innovation, and the broader integrity of digital markets. This multi-statute environment generates persistent uncertainty. Key unresolved issues include the permissibility of collecting data that is publicly visible on websites, the extent to which platforms exercise de facto control rights over information they host, the legal significance of bypassing anti-crawling measures, and the degree to which technological neutrality should shape judicial reasoning. As a result, automated collection becomes not merely a technical operation but a legally contested practice embedded within China's broader data governance landscape.

By contrast, Singapore follows a sharply different model. Its Computer Misuse Act, Personal Data Protection Act, and the Copyright Act's explicit exception for text and data mining activities form a vertically integrated regulatory system [3]. This structure is centered on authorization, system integrity, and stringent controls on personal data. It provides clearer compliance boundaries and more predictable legal outcomes than China's decentralized framework. Because China and Singapore are both prominent digital economies in Asia yet rely on fundamentally different legal philosophies, comparing the two systems offers a valuable analytic lens through which to examine how web crawling ought to be governed in an era defined by artificial intelligence and large scale computational analysis.

1.2 Research Gap: Absence of Empirical Evidence and Absence of Cross-Jurisdictional Comparison

Most existing scholarship addresses doctrinal debates within China or focuses on theoretical questions involving data rights, platform governance, and privacy. Very few studies rely on comprehensive empirical evidence to show how Chinese courts adjudicate disputes involving automated data collection. Even fewer situate China's practice in a comparative context. This absence is striking because Singapore is the only Asian jurisdiction that simultaneously criminalizes unauthorized system access through a dedicated statutory regime, regulates personal data collection through a comprehensive privacy framework, and formally recognizes text and data mining for research and computational analysis. A systematic comparison between China and Singapore can reveal deep differences in regulatory philosophy and can clarify how Asian jurisdictions conceptualize crawling in ways that reflect their respective priorities in innovation, security, and economic governance.

This study responds to two major research gaps. It addresses the lack of empirical analysis concerning Chinese judicial practice and the lack of cross jurisdictional comparison that situates China's evolving model alongside Singapore's more consolidated and vertically structured regime. Together, these gaps limit academic understanding of how the law functions in practice and obscure how Asian states articulate divergent visions for the management of automated data collection.

1.3 Data Source and Contributions of This Study

To analyze judicial practice in China, this study constructed a custom dataset obtained through automated extraction of judgments published on China Judgments Online. The dataset spans a ten year period and includes several hundred decisions involving disputes such as unfair competition,

copyright infringement, personal information processing, contractual conflicts surrounding application programming interface access, and criminal prosecutions for unauthorized system entry. Each judgment was transformed into structured variables capturing legal basis, defendant type, procedural posture, judicial outcomes, defense strategies, geographic distribution, and related attributes. A diverse set of analytical diagrams, including dashboards, Pareto distributions, correlation matrices, network visualizations, temporal trend graphs, defense success rates, and legal basis frequency charts, supports the empirical analysis and reveals structural patterns within judicial reasoning. These visual materials illustrate the configuration of the dataset and enable cross variable interpretation.

In addition to its empirical contribution, the study advances a comparative analysis of Singapore's access centered regulatory system. The juxtaposition of China's fragmented, market oriented model with Singapore's unified, authorization based approach illuminates the conceptual differences underlying the two jurisdictions. The combined empirical and comparative framework allows this research to demonstrate not only how Chinese courts adjudicate crawling disputes in practice but also how China's evolving system fits into a wider Asian landscape where data governance models diverge sharply.

2. Methodology

This study is based on an independently constructed dataset of judicial decisions collected from China Judgments Online, the official national platform for court rulings. Due to anti-automation measures and dynamic page structures, manual collection was impractical. A custom acquisition system was therefore developed using Selenium, undetected chromedriver, pandas and openpyxl to simulate human browsing behavior and ensure stable data extraction. The dataset covers the period from 2013 to 2023 and includes information on case type, court level, region, judicial reasoning, judgment outcomes and applied laws.

A strict filtering process was applied to retain only cases in which automated data collection constituted a substantive part of the dispute. Keyword-based screening was used to identify crawler-related cases, while cases in which online activity appeared only as background context were excluded. The final dataset contains more than five hundred civil, administrative and criminal cases.

To enable systematic analysis, metadata inconsistencies across jurisdictions were standardized. Case types and judgment outcomes were unified, geographic information was aligned with official court classifications, and defence arguments were coded into structured categories. Legal bases were extracted through automated parsing and manually verified. Descriptive visualisations were used to assess data structure and representativeness, while statistical relationships between key variables were examined using correlation analysis. All data processing and visualization were conducted in Python to ensure transparency and reproducibility.

3. Results

3.1 Structural Characteristics of Crawler-Related Litigation

The empirical observations portray a complex and evolving judicial landscape surrounding automated data collection in China. Civil cases constitute the overwhelming majority of disputes, consistently accounting for more than eighty percent of the dataset, while criminal and administrative cases remain comparatively limited. This distribution indicates that Chinese courts predominantly conceptualize web crawling as a commercial and competitive practice rather than conduct associated with cybercrime or administrative misconduct.

The regional distribution of disputes further illustrates this structural pattern. As shown in Figure

1, crawler-related cases are heavily concentrated in economically advanced regions, including Beijing, Guangdong and Shanghai. This concentration closely corresponds with the clustering of major platform companies and data-intensive industries, highlighting the role of digital platform economies in generating conflicts related to automated data extraction.

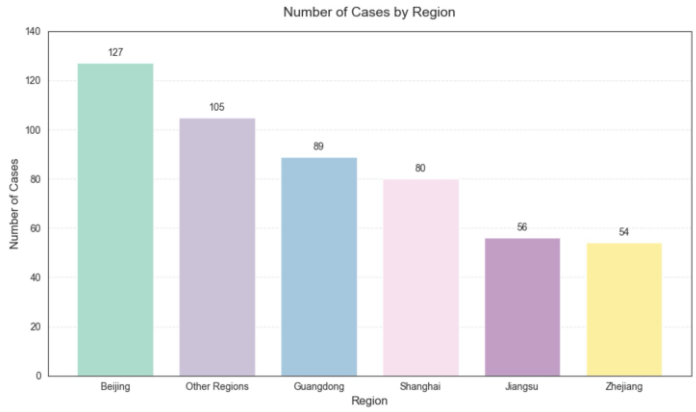

Figure 1: Regional Distribution of Cases.

The distribution of defendant types reinforces the economic nature of these disputes. Figure 2 illustrates that enterprises constitute the majority of defendants, while individuals and public institutions appear far less frequently. This asymmetrical pattern suggests that legal risks linked to automated extraction fall disproportionately on commercial entities, which face heightened scrutiny for potentially disrupting market order or platform stability[4].

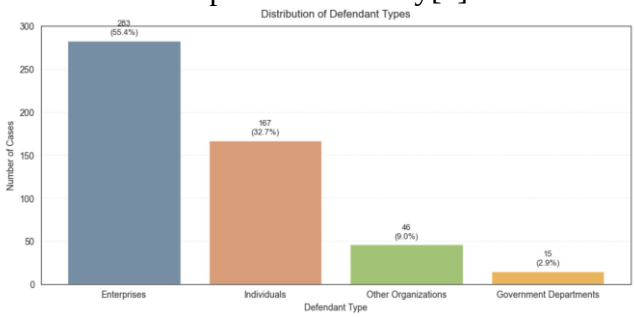

Figure 2: Defendant Type Distribution.

3.2 Doctrinal Foundations and Legal Basis

Analysis of case reasons reveals the doctrinal foundations guiding judicial intervention. The most prevalent categories involve disputes concerning infringement of the information network dissemination right and unfair competition. This distribution reflects a dual doctrinal framework.

The statutory structure underlying judicial reasoning is further illustrated by citation patterns. As shown in Figure 3 courts most frequently rely on the Civil Code, Copyright Law and Anti-Unfair Competition Law. Where automated extraction involves copyrighted images, audiovisual materials or text, courts tend to rely on copyright doctrine[5]. In contrast, when extraction targets publicly accessible but platform-controlled data—such as product listings, pricing information or user-generated content—courts more often apply competition law principles. Together, these patterns demonstrate that Chinese courts rely on a multi-layered interpretive structure rather than a single statute when evaluating the legality of automated data acquisition [6].

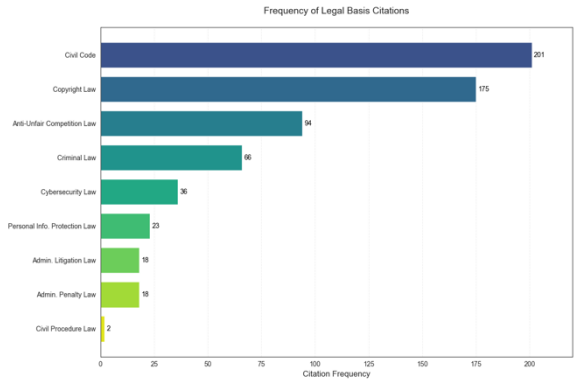

Figure 3: Legal Basis Citation Frequency.

Further insight into the structure of crawler-related disputes is provided by the distribution of causes of action. As shown in Figure 4, a small number of dispute categories account for the majority of cases, indicating a highly concentrated pattern of legal conflict. Claims based on infringement of the information network dissemination right and unfair competition dominate the dataset, while other causes of action appear far less frequently.

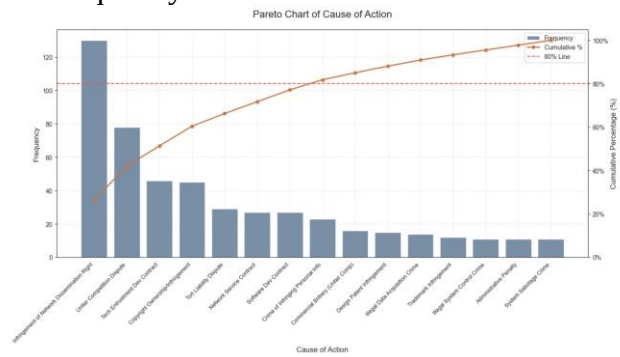

Figure 4: Pareto Distribution of Case Reasons.

This concentration suggests that judicial scrutiny of automated data collection consistently centers on a limited set of core legal concerns. Rather than addressing a wide range of fragmented legal issues, courts repeatedly evaluate crawling behavior through recurring doctrinal lenses related to content dissemination, market competition, and access control. The Pareto distribution therefore highlights the structural regularity of crawler-related litigation and reinforces the view that automated data extraction disputes are shaped by a stable and identifiable set of legal risk categories.

3.3 Judgment Outcomes and Temporal Evolution

Judgment outcome patterns reveal a nuanced judicial approach. Partial support rulings constitute the most common outcome, exceeding both full support decisions and dismissals. This indicates that courts rarely accept plaintiffs' claims in their entirety and instead calibrate remedies, considering the legitimate role that automated extraction plays within the digital ecosystem.

The temporal evolution of judicial outcomes further clarifies this trend. Figure 5 illustrates the distribution of civil judgment outcomes between 2013 and 2022. Partial support rulings show a gradual increase over time, while full support and dismissal outcomes remain relatively stable. This pattern suggests growing judicial sophistication in balancing technological considerations with economic and regulatory concerns.

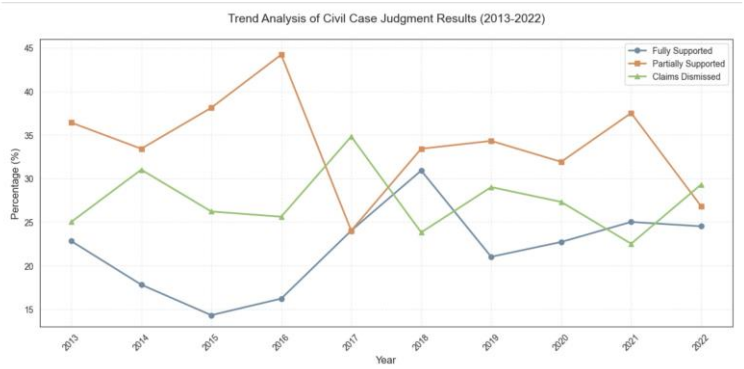

Figure 5: Temporal Evolution of Civil Judgment Outcomes.

3.4 Plaintiff Claims and Judicial Outcomes

The relationship between plaintiff claims and judicial outcomes provides additional insight into adjudicatory logic. As shown in Figure 6, different categories of plaintiff claims correspond to distinct judgment outcomes. Claims grounded in clearly defined legal interests, such as contractual rights or identifiable data infringement, are more likely to receive partial support, whereas broadly framed or weakly substantiated claims more often result in dismissal. This distribution indicates that courts place significant weight on claim specificity and legal grounding when assessing crawler-related disputes.

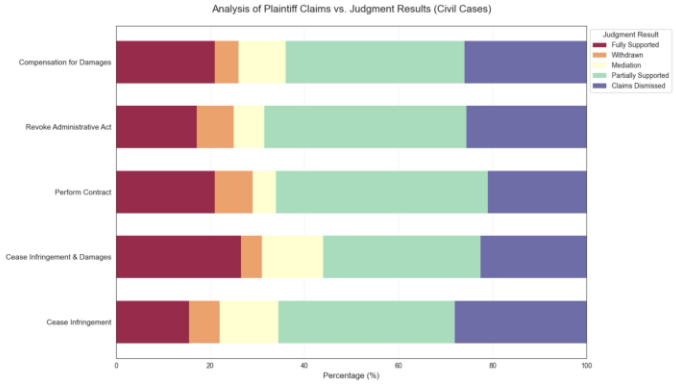

Figure 6: Flow of Plaintiff Claims to Judicial Outcomes.

Taken together, the results depict a context-dependent and balancing-oriented judicial approach to automated data collection. Chinese courts neither categorically prohibit nor broadly endorse web crawling. Instead, legality is assessed through a combination of doctrinal analysis, market considerations and technical context. These empirical patterns provide the foundation for the comparative analysis that follows.

4. Discussion

The empirical findings reveal how Chinese courts conceptualize the legality of automated data collection through a layered and context-sensitive framework. Rather than treating web crawling as a uniform technological behavior, courts interpret it as an economic and legal practice shaped by market competition, data access arrangements, private ordering mechanisms, and the country's evolving data governance regime. At the core of this approach lies a persistent tension between platform control and data accessibility, which reflects broader debates over the value and permissible use of data in the digital economy.

A central feature of the judicial landscape is the dominance of civil litigation. Courts overwhelmingly frame crawling disputes as conflicts between economic actors rather than matters of public order or criminal wrongdoing. Even in cases involving large-scale extraction or the circumvention of technical measures, criminal sanctions remain relatively rare. This pattern suggests that Chinese courts view web crawling as a technologically neutral tool whose legality depends on context, particularly the competitive relationship between the crawler and the platform[9]. Litigation arising from e-commerce, digital media, and data-driven services further supports the conclusion that crawling disputes are primarily evaluated through an economic lens.

The prominence of enterprise defendants reinforces this interpretation. Commercial actors typically deploy crawlers to aggregate data, monitor competitors, or develop derivative services, which places their conduct under closer judicial scrutiny. Courts frequently rely on the Anti-Unfair Competition Law to assess whether such practices disrupt market order or unfairly exploit platform resources [10]. By contrast, individuals are less frequently targeted and usually appear in cases involving limited or low-impact extraction, indicating a differentiated judicial response based on scale and market impact.Judicial reasoning also reveals a dual doctrinal structure. When automated extraction involves copyrighted images, videos, or textual content, courts tend to apply copyright principles. Where extraction targets publicly accessible but platform-controlled data, competition law and contractual doctrines play a more prominent role. This multi-layered approach demonstrates that courts do not rely on a single statute but instead draw on multiple legal sources to resolve technologically complex disputes[7].

Technical protection measures have become an increasingly important factor in judicial assessment. Courts often treat anti-crawling technologies, login requirements, rate limiting, and robots protocol enforcement as signals of a platform's intention to define access boundaries. The circumvention of such measures is frequently characterized as a violation of good faith and a disruption of competitive balance. At the same time, courts remain cautious not to equate all automated access with illegality, particularly where extraction does not interfere with platform operations or harm user interests.The high proportion of partial support judgments reflects this balancing approach. Courts regularly limit overreaching claims by platforms that seek exclusive control over publicly accessible data. Defences based on lawful use, public data, or existing contractual authorization achieve relatively higher success rates, indicating judicial recognition that not all crawling activities are harmful. Correlation and network analyses further show that courts anchor their decisions in statutory interpretation rather than purely technical evaluation, with legal basis functioning as the central element connecting case type, defence strategy, and judgment outcome[8].

Recent trends also reveal the growing influence of data governance considerations. Increasing citations of the Cybersecurity Law and the Personal Information Protection Law indicate heightened judicial attention to system security and personal data protection [11]. This shift reflects China's broader regulatory turn toward data sovereignty and personal information governance, even as courts continue to accommodate legitimate technological and economic uses of automated access.

Taken together, the discussion suggests that Chinese courts neither categorically prohibit nor fully endorse web crawling. Instead, they adopt a flexible and context-dependent approach that balances innovation, competition, platform stability, and data governance. While this interpretive flexibility allows courts to respond to novel disputes, the absence of a unified legal framework for data access generates uncertainty for developers, researchers, and firms that rely on automated data collection. The comparative insights drawn from Singapore highlight potential directions for reform. Clearer access rules and explicit provisions for text and data mining could reduce uncertainty, while China's competition-based reasoning offers a useful counterbalance against overly restrictive access regimes. A more coherent governance framework that integrates access clarity with proportional regulation

would better align legal practice with the realities of data-driven innovation.

References

[1] Khder, M. (2021) Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. International Journal of Advances in Soft Computing and Its Applications. 13, 145–168.

[2] Razzaq, A. and Yang, X. (2023) Digital finance and green growth in China: Appraising inclusive digital finance using web crawler technology and big data. Technological Forecasting and Social Change. 188, 122262.

[3] Fiil-Flynn, S.M., Butler, B., Carroll, M., Cohen-Sasson, O., Craig, C., Guibault, L., et al. (2022) Legal reform to enhance global text and data mining research. Science. 378, 951–953.

[4] Aejas, B., Belhi, A., and Bouras, A. (2025) Using AI to Ensure Reliable Supply Chains: Legal Relation Extraction for Sustainable and Transparent Contract Automation. Sustainability. 17, 4215.

[5] Pan, J. (2025) Research on the Legal Regulation of New Unfair Competition Behavior of Network Platform. Scientific Journal Of Humanities and Social Sciences. 7, 93–101.

[6] Liu, X. (2022) China's Anti-Unfair Competition Law on Data Crawling: in: Dalian, China.

[7] Leitao Requena, D. (2025) How and Why Organizations Litigate: Empirical Evidence from Dutch Commercial Court Cases, dr., Utrecht University, 2025.

[8] Yu, X. (2022) The three legal dimensions of China's big data governance. Journal of Chinese Governance. 7, 511–530.

[9] Chen, J. (2025) Empirical Study on the Regulation of Data Crawling Behavior under the Anti-Unfair Competition Law. Advances in Social Behavior Research, 16(2), 1–11.

[10] Gu, J. (2023) An Empirical Study on the Judicial Regulation of Data Crawling Unfair Competition. International Journal of Education and Humanities. 9, 61–66.

[11] Calzada, I. (2022) Citizens' Data Privacy in China: The State of the Art of the Personal Information Protection Law (PIPL). Smart Cities. 5, 1129–1150.