When web crawlers infringe personal information: Judicial evidence, legal governance and legalistic swamp of China

Yangkun Huang[a] and Sini Su[b]*

[a] *School of Media and Communication, Shanghai Jiao Tong University, Shanghai, China;* [b] *College of Media and International Culture, Zhe Jiang University, Hangzhou, China.*

Correspondence: 866 Yuhangtang Road, Xihu District, Hangzhou, 310058, 21923038@zju.edu.cn.

When web crawlers infringe personal information: Judicial evidence, legal governance and legalistic swamp of China

Web crawlers, coupled with the advent of the big data era, have been widely recognized and utilized for the capacity to capture massive resources on the internet. Nonetheless, it is in the process of technology application and promotion that crawler technology is abused and misused, and then malicious crawlers have been dragged into public view, posing a threat to netizens' data privacy. Based on 103 legal documents disclosed to the Chinese public, this study conducts the typological coding of factors including the criminal subject and object, the criminal acts, and penalty with fine, and evaluates the judicial application through data analytics. This research provides an overview of Chinese judicial practices in cracking down on illegal activities of infringement upon personal information via web crawler, with efforts to assess the applicability, effect, and shortcomings of Chinese legal governance on this issue.

Keywords: Web crawler; Personal information; Legal documents; Typological analysis

1. Introduction

To fully enjoy cyberspace and cloud environment, people tend to hand over the reins of *personal information (PI)* to a third party, including government, institution, and other online service providers (Jahankhani, Al-Nemrat, & Hosseinian, 2014). Nonetheless, organizational causes, e.g., database security vulnerabilities, and governmental causes, e.g., online governance weakness, would subsequently incur the potential malicious attack, jeopardizing netizens' personal data. One of the biggest dangers, there among, lies in web crawlers. A report showed that, when surfing the internet, we are often immersed in the digital community but kept in the dark about web crawlers: in 2020, malicious web crawlers, so-called bad bots, comprised 25.6% of the web traffic, reaching the highest proportion ever (Imperva, 2021).

In China, big data technology developing prosperously with web crawlers used widely, bad bots have undoubtedly become a present danger to citizens' data privacy. Although individual data literacy could not be overemphasized in nowadays data-driven society (Pangrazio, & Sefton-Green, 2020), fostering multi-entity engagement at every level, individuals, organizations, and regulators, of anti-crawler mechanisms in China will be no easy feat — For one thing, due to the virtuality and anonymity of the cyberspace, internet users can barely detect automatic vectors in stealing personally identifiable information, let alone guard against possible infringements caused by crawlers. For another, crawler-related criminal cases with great influence mainly involved enterprises and organizations, such as the case of Baidu vs. Dianping's unfair competition (Baidu Co., Ltd., v. Jietu Co., Ltd., 2016) and the case of Shanghai Shengpin Co., Ltd. illegally obtaining computer information system data (People's Procuratorate of Haidian District v. Shengpin Co., Ltd., 2017), while media and press shed disproportionate lights on individuals at risk in a virtual world marked by bad bots. The role of confronting malicious crawlers, thus, is mostly given to the internet

governors, and the question to ponder upon is left for them that how to manage the governance of the crawler technology and protect the *PI* from bad bots against the backdrop of the ever-growing internet population.

China welcomed its billionth netizen as of June 31, 2021 (China Internet Network Information Center, 2021), where increasing population size purports the online prosperity of *PI*, as well as the mounting challenges for the national legal governance systems: There have been a bunch of criminal cases of *infringing upon citizens' personal information (IUCPI)* via web crawler in China, one of which embroiled as many as 8.3 billion pieces of *PI* from social engineering database (People's Court of Huai'an v. Niu, 2018). With malicious crawlers dragged into public view, beneficial and essential is providing an overview of the conviction and sentences imposed by Chinese courts in prosecutions of invading personal data via web crawlers according to existing judicial decisions, which would contribute to identifying the status quo, namely the governance model and bottlenecks, of the punishment on abuse and misuse of crawler technology *to* harm personal data security, and can also offer Chinses insights into legal governance on the relationship between the web crawler and the *PI* for global practices.

2. Background Literature

2.1 Personal information protection: Why data matters and How governors (should) secure data

We are surely ushering in a new phase of digitization: Other than providing personal demographic information such as marital status, education, occupation, and age (Phelps, Nowak, & Ferrell, 2000), many of us are, voluntarily or otherwise, engaged in the quantified-self movement where our biological, physical, behavioral, and environmental information on different aspects of the daily lives has been collected with the aid of invisible technology (Swan, 2013; Marcengo, & Rapp, 2014). Undeniable as the benefits in our provision of personal data to third parties are, for instance, personalized services like news and product recommendation could be customized by ourselves according to data offered, where there are interests and benefits, there are disputes and chases, both legitimate and illegitimate: Data brokers are everywhere, collecting, packaging and selling your *PI* sometimes at the margins of the law (Brooks, 2001; Otto, Antón, & Baumer, 2007; Yeh, 2018).

The question, therefore, comes down to why personal data matters critically. In fact, many countries initially protected *PI* from the perspective of a right of personality, namely, the right of privacy, drawing a line between the information privacy as personhood and other personal possessions as properties, and China is no exception (Whitman, 2003; Prins, 2006; Xie, 2015; Lv, 2019). In this view, *PI* essentially has no property attributes in the legal sense, or can be regarded as a kind of humanistic property, which means that there should not have been legal controversy on the economic aspects of personal data privacy. The situation, however, has changed that the distinction has become blurred, especially when the politics and business giants have realized that personal data accumulation in the network, which can be mined through data analytics in constant progress, does shape a useful second-self of each individual, playing a transformative role in political campaigns (Dommett, 2019), public administration (Fahey, & Hino, 2020), and electronic commerce (Acquisti, Taylor, & Wagman, 2016). Emerging forms of commodification and utilization of personal data echo the perspective that data privacy can be cast as a property right (Murphy, 1995; Schwartz, 2003; Chellappa, & Shivendu, 2007; Liu, 2007; Varian, 2009; Hong, & Jiang, 2019). As such, *PI* is recognized for its monetary value, which, plus *PI*'s natural attribute of personhood, explains why we should lay stress on the importance of personal information protection.

Some scholars propose that the existing governance framework for personal information protection, which was born in the pre age of big data, has continued to be severely challenged and is no longer tenable now for the reason that the sociality is to replace individuality as the main attribute of *PI* (Wang, & Chang, 2020). And as already mentioned, the role of protecting *PI* is usually given to the governor itself. Among all tools, laws, with legal governance, are undoubtedly directly binding. An overview of lawmakers from different backgrounds reveals that legal frameworks vary among countries: The EU *General Data Protection Regulation (GDPR)* is obviously in line with the viewpoint of Westin (1967), demonstrating that all citizens have the sole right to control over their personal data, including rights to access data, rectification, erasure, and object. In essence, the *GDPR* empowers the citizens as users with enforceable rights (Tene et al., 2019). While different from the EU, the US does not express the idea of *PI* protection in a harmonized legislation; rather, current legislation of the US covering the *PI* protection is dispersed in various laws on both federal and state levels, and in the US, governance differs from state to state. Another distinction is that the US has passed a pile of sector-specific laws constraining *PI* usage in industries like healthcare and telecommunication, and accordingly, industry self-disciplines are combined with these laws (Pittman, & Levnberg, 2021). As for the legal governance practice in the eastern world, China's *PI* protection has recently moved into a new stage where *Personal Information Protection Law (PIPL)* has been formally passed into law and is going to come into effect on November 1, 2021, addressing gaps that there is no specialized law on *PI* before. Borrowing a lot from the *GDPR*, the *PIPL* shares the idea of citizen empowerment, meanwhile underlining the *PI* protection in the context of national and public security (Personal Information Protection Law of the PRC, 2021).

There is indeed a governor-level awakening sense of personal information protection, and a number of countries have already acted to strengthen legal controls. Yet the picture of the *PI* infringement is far more complex; for example, netizens, particularly when faced with harm through the invisible background codes such as algorithm and crawler, would be more likely to go through difficulties in detecting and reporting that their *PI* is under threat, not to mention urging rights-based data governance. Does legal governance make a difference in such circumstances? If it is true, how does it work? Apparently, detailed matters are pending further clarification.

2.2 Grey zone in digital era: Legal issues associated with web crawlers

Nowadays, development, no matter for individuals, organizations, or society, is inseparable from data. As a result, the web crawler emerged as a means to collect internet information efficiently. Technically, a web crawler, so-called web spider or web scraper, is a string of codes or a program developed for automatically browsing and downloading structural web contents. Crawler technology was initially used in search engines to index web pages for users, and with the continuous maturity of the concept and practice of big data, it now functions as the cornerstone of the open Internet and shared e-resource, outreaching to more fields (Cui, & Xu, 2019). It is, nevertheless, in the process of technology application and promotion that crawler technology is abused and misused, and malicious crawlers sprang up, doing evil, including content and price scraping, account takeover and creation, and credit card fraud (Imperva, 2020).

Actually, legal issues have centered around this technology since the birth of web crawlers. Some Chinese law practitioners have been keenly aware of the crawling technology, and according to China's current laws, summarized potential risks of violation of laws and regulation in the course of using crawling technology (Wang, & Chen, 2019a, 2019b), which consist of:

1) the unfair competition;

2) the infringement upon the right of dissemination over information networks;

3) the crime of infringing upon citizens' personal information *(IUCPI)*;

4) the crime of illegally obtaining data from the computer information system *(IODCS)*;

5) the crime of illegal intrusion into the computer information system *(ICS)*;

6) the crime of providing programs and tools for invading and illegally controlling the computer information system *(PTICCS)*.

Compared with standpoints of practitioners, the focus of legal scholars goes beyond tort and crime. An intriguing fact is that Chinese researchers seem to be much more attracted to this topic. And early-stage attention has paid to a marginal issue — the legal force of the robot exclusion protocol, somehow a gentleman's agreement, and a lot of discussions have been dedicated to whether the protocol can constrain unfair competition among internet companies (Zhang, 2013; Yang, 2014; Ning, & Wang, 2016). While the more profound the understanding of crawlers' legal threats, the more comprehensive the research perspective. Then scholarly concern finally turned to the tool usage itself: Professions have already systematically considered the legal boundaries for crawler-tool usage (Gold, & Latonero, 2017; Li, & Sun, 2018; Yang, 2020; Zhang, 2020; Krotov, Johnson, & Silva, 2020; Ruan, 2021; Su, 2021), and almost all of consideration holds the basic principle of technological neutrality that the crawler technology is a double-edged sword, depending on the user's intention. Simultaneously, based on these fundamental combing and discussing, views of some others have been extended to some specific crimes in Chinese criminal laws, e.g., the crime of *IODCS* (You, & Ji, 2019) and the crime of infringing upon intellectual property (Xu, 2020). Among them, what has received the most attention is the use of crawlers in *IUCPI* (Liu, 2019; Xu, & Zhang, 2020; Song, 2021), mirroring that the law community has set about to contributing their insights into addressing the relationship between a natural tool and a core right.

Very regrettably, enthusiastic though Chinese researchers are, and abundant though relevant criminal cases, there are few studies built on empirical document analysis, and almost all of them are conducted with critical and imaginary thinking, which is from where the research starts.

The research focuses on the typological features of the suspects, the information infringed, and the conviction and sentencing of the crimes of *IUCPI* via web crawler and aims to assess the following questions:

1) What basic features are there in criminal cases of *IUCPI* via web crawler?

2) What basic features are there in sentenced defendants in the crime of *IUCPI* via web crawler?

3) How are the relevant provisions of the crime of *IUCPI* expressed in the application of current Chinese laws?

4) What is known about the governance experience and judicial dilemma in legal practices of the crime of *IUCPI* via web crawler?

3. Materials and Methodology

3.1 Data

Available full-text legal documents for this study are retrieved from the China Judgements Online *(JOC)* website, the globally largest open and continuously updated database for judicial documents, and also the official website for effective judgments of Chinese People's Courts at all levels (Luo, 2020). Due to the massive amount of legal documents, the *JOC* website has been a preferred research source for legal issues in China (Zhang, 2016; Gao, & Wen, 2017; Zhou, Peng, & Bao, 2017; Xia et al., 2020).

Specifically speaking, all materials are collected with the Boolean retrieval via the alternate

keyword for the cause of action, each in combination with the keyword "web crawler" and its synonyms for full-text searching. Table 1 displays keywords used for document collection. Noteworthy in terms of the cause of action is that, in addition to the crime of *IUCPI*, there are two other crimes selected. It is because the crime of *IUCPI* is the combination of these two crimes, the crime of *SPCPI* and the crime of *IOCPI*, both of which have been abolished since 2015.

Table 1 Keyword list for Boolean retrieval

Element	Keyword
Cause of action	Crime of infringing upon citizens' personal information *(IUCPI)*
	(Qīnfàn gōngmín gèrén xìnxī zuì)
	Crime of selling or illegally providing citizens' personal information *(SPCPI)*
	(Chūshòu, fēifǎ tígōng gōngmín gèrén xìnxī zuì)
	Crime of illegally obtaining citizens' personal information *(IOCPI)*
	(Fēifǎ huòqǔ gōngmín gèrén xìnxī zuì)
Criminal tool	Web crawler (Páchóng)
	Web spider (Wǎngluò zhīzhū)
	Web spiderbot (Wǎngluò jīqìrén)
	Web scraper (Cǎijí qì)
	Web crawler script (Jiǎoběn)
Modus operandi	Illegally crawling (Zhuā qǔ)
	Illegally scraping (Pá qǔ / Bā qǔ)
	Illegally collecting (Cǎijí)

The date for the first collection time is June 1st, 2021. And in order to keep the research up to date, the researchers conducted the second collection on October 1st, 2021. Finally, 103 judges from 2015 to 2021 remain after the removal of duplicated and extraneous materials.

3.2 Methods

The methodology that is central in this research, is dealing with the specific legal facts that can be converted into standardized information in a systematic and typological way, and the dataset is comprised of 103 cases and 288 offenders with the other seven defendant units excluded, is analyzed quantitatively based on the method of document analysis in the criminology (Epstein, & Martin, 2010).

Four aspects, namely, case, offender, fact, and outcome, are extracted from the written judgment documents. More concretely, encoded variables are shown in Table 2. To assess differences, statistical analyses, including descriptive statistics, nonparametric tests, were conducted using SPSS 24.0.

Table 2 Coding scheme

Category	Variable
Case	Year
	Place
Offender	Age
	Gender
	Place of origin
	Level of education
	Occupation
Fact	Type and amount of the *PI* infringed

	Amount of money involved
	Type of illegal act
	Severity of the circumstances
Outcome	Type and length of sentence
	Amount of fine

4. Results

4.1 Overall Pattern of Cases: Descriptive statistics of the criminal cases of *IUCPI* via crawlers

The distribution of criminal cases of *IUCPI* via web crawlers shows some disparity by year and region (Figure 1): the number of criminal proceedings peaked in 2018, a total of 40 cases. In comparison, the crawler's illegal damage to the *PI* occurred much less often during the years before and after 2018. On the other hand, the regional distribution was characterized by an obvious imbalance to the extent that the total quantity of cases in Central *(N=24)* and Western *(N=17)* China is far less than that in the East *(N=62)*.

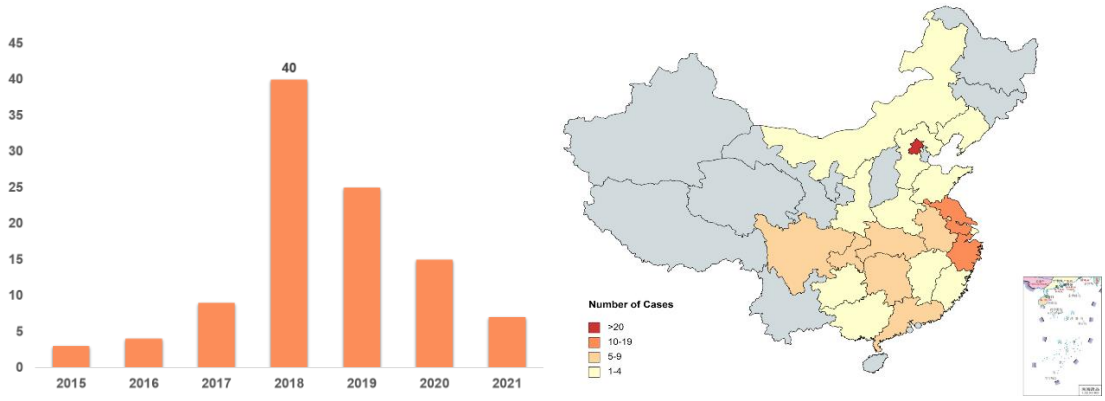

Figure 1 Yearly and regional distribution of cases

Such spatial-temporal characteristics do not come as a surprise: The official interpretation on criminal cases involving *IUCPI* went into effect before 2018, identifying clear standards and providing reliable references for conviction and sentencing, and the surge in the cases number in 2018 can be partly attributed to the clarity and disambiguation of law, while judicial efforts made to improve the information ecosystem governance, in turn, conveyed the idea that *PI*-stealers have been and will be prosecuted, to some degree, deterring the crime in the social sphere. And concerning that more than 60% of cases are in Eastern China, the feature corresponds to the development of the information industry in these provinces. After all, one of the main motivations for using crawler tools is to obtain network information quickly, out of strong demands for data.

In Table 3, descriptive statistics about the socio-demographic information of defendants are reported. Considering the incomplete documentation, researchers only the identifiable data is counted for the calculation of the distribution among factors. As a whole, peculiarities of 288 criminals' portraits are concluded: middle-aged, male, geographically dispersed, high-educated, and unemployed.

Court outcomes are presented in the bottom section of Table 3. In terms of court outcomes, acts of 129 offenders' *IUCPI* were identified as "serious circumstances" while the other 159 involved "especially serious circumstances". But according to Criminal Law of the PRC, where the circumstances are "especially serious", personnel involved in activities of *IUCPI* shall be sentenced to fixed-term imprisonment of not less than three years but not more than seven years with fine (Criminal Law of the PRC, 2020). Thus it is obvious that the Chinese judicial system used to have

a lighter sentence for a 3rd Quantile of 38 months for the sentence length. Additionally, it can be seen that in China, levying penalties is a general practice, producing punishment and prevention effects on profit-making crimes like the crime of *IUCPI.*

Table 3 Descriptive statistics of the offenders and court outcomes

	Total
Age group	
＜18	0 (0.00%)
18-29	57 (24.57%)
30-39	144 (62.07%)
40-49	30 (12.93%)
50 above	1 (0.43%)
Unidentified	56
Gender	
Male	257 (93.80%)
Female	17 (6.20%)
Unidentified	14
Place of origin	
Eastern China	97 (36.19%)
Central China	101 (37.69%)
Western China	70 (26.12%)
Unidentified	20
Level of education	
Illiterate	0 (0.00%)
Primary	14 (5.57%)
Junior secondary	52 (20.72%)
Senior secondary	65 (25.90%)
Higher	120 (47.81%)
Unidentified	37
Occupation	
Company staff	72 (33.18%)
Farmer	14 (6.45%)
Merchant	4 (1.84%)
Public servant	3 (1.38%)
Worker	11 (5.07%)
Unemployed	113 (52.08%)
Unidentified	71
Severity of the circumstances	
Serious	129
Especially serious	159
Sentence type	
Fixed-term with penalty	282
Only imprisonment	0
Only penalty	5
Exemption	1

Sentence length	
1st Quantile	12
Median	24
3rd Quantile	38
Max	60
Fine in RMB	
1st Quantile	8,000
Median	15,000
3rd Quantile	40,000
Max	1,100,000

The result of our encoding of legal facts shows (See Table 4) that telecom contact, i.e., phone and cellphone number, through which people can be reached directly although identified vaguely, is the most frequently-infringed *PI* type, and 73 out of 103 cases are related to its illegal theft. In regard to the amount of the *PI* infringed, the median of 71317, reaching the highest sentencing standard, demonstrates that a vast amount of *PI* is now put in a vulnerable situation where spider bots run wild, while the amount of money involved in these crawler cases is not as large as expected, perhaps because bad bots were under surveillance before the *PI* was traded for unlawful economic benefits.

Besides, warning signals are worth noting that at present, more than half of criminals have already participated in multi-stages of *IUCPI*, from tool development, data crawling, to data exchange and sales, even to follow-up illegal activities, including lending (People's Procuratorate of Yuanjiang City v. Li, 2020), fraud (People's Procuratorate of Tiandong County v. Mo, 2018), and selling fraudulent medicines (People's Procuratorate of Chen'an County v. Gong, 2018).

Table 4 Descriptive statistics of the legal facts

	Total
Type of the *PI* infringed	
Name	39 (37.86%)
ID card	19 (18.45%)
Telecom contact	73 (70.87%)
Address	38 (36.89%)
Account password	53 (51.46%)
Property status	14 (13.59%)
Personal whereabouts	3 (2.91%)
Amount of the *PI* infringed	
1st Quantile	10,000
Median	71,317
3rd Quantile	1,010,000
Max	8,435,371,763
Amount of money involved in RMB	
1st Quantile	11,970
Median	37,000
3rd Quantile	115,000
Max	1,000,000
Type of illegal act	

Only illegally obtaining (OIO)	89 (30.90%)
Only selling or illegally providing (OSP)	45 (15.63%)
Multi-stage engagement (MSE)	154 (53.47%)

4.2 Assessing the application of current law: Nonparametric tests on differences in sentencing

Descriptive statistics above offer an overview of the conviction and sentences imposed by Chinese courts in prosecutions of invading personal data via web crawlers. When it comes to, however, that the current law concerning the *IUCPI* crime is applicable to Chinese realities in the Big data age, more relationships need to be clarified.

A range of statistical tests suggests that in a general sense, the dilemma of application of current law in China is not that evident on the matter of illegal crawler usage for infringing upon Chinese citizens' personal information. Nonparametric tests were performed to reveal differences in sentencing and conviction across cases of varying scale and severity. In this study, the classification of the severity of cases, two types as serious circumstance and especially serious circumstances, is in line with the PRC criminal law, And the volume of the infringed citizens' *PI*, as well as the amount of money involved, were divided into four levels based on quartile calculations.

As expected, results of the Mann–Whitney U tests indicate that there are significant differences both in the sentence length and fine between the two circumstances. Those who are suspected of being involved in especially serious circumstances will face higher fines (median = 30,000, n= 159) and longer prison sentences (median = 36, n= 159). Box plots (See Figure 2) below show these differences more visually.

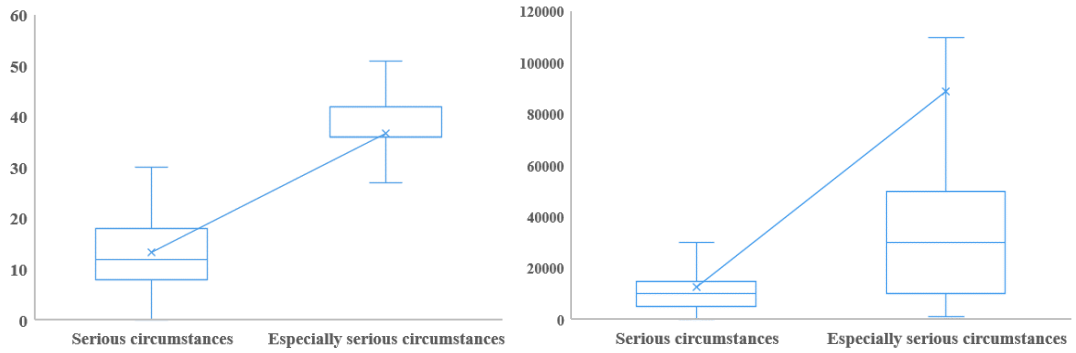

Figure 2 Box plots of sentence length and fine between two circumstances

It is perfectly logical that the Kruskal-Wallis tests also reveal significant differences in the sentence length and fine across four levels of the amount of the *PI* infringed (H_{st}= 83.172, p<0.001; H_f= 34.291, p<0.001;), and differences are the same significant across levels of the money involved (H_{st}= 15.268 p=0.002<0.01; H_f = 58.688, p<0.001). In fact, PRC Criminal Law does not specify the conviction standards for the amount of information and money, but relevant provisions are expressed in a judicial interpretation, which in China, has a universal judicial binding effect too (Interpretation of the Supreme People's Court and the Supreme People's Procuratorate on Several Issues concerning the Application of Law in the Handling of Criminal Cases of Infringing on Citizens' Personal Information, 2017).

Nevertheless, our concerns for the *PI*, based on the characteristics of the era and technology, are not groundless. Dunn tests, a post hoc test, were followed to determine the nuances after the discovery of significant differences.

Dunn's t-test pairwise comparison tells that variations in terms of imprisonment and forfeit tend to be not evident among extreme cases where not so much, or too much, *PI* was involved.

According to Table 5, there is no significant differences in the sentence length and the fine between the first level, less than 10,000 pieces of *PI*, and second level, more than 10,000 but less than 71,317 pieces; And the condition is all the same between the third level, more than 71,317 but less than 1,010,000 pieces of *PI*, and fourth level, over 1,010,000 pieces.

Table 5 Dunn's t test pairwise comparison on sentence length and fine for amount of the *PI* infringed

	(I) Amount of the PI infringed	(J) Amount of the PI infringed	(I) Median	(J) Median	Difference	*p*
Sentence length	1st level	2nd level	14	16	-2	0.178
	1st level	3rd level	14	36	-22	0.000**
	1st level	4th level	14	37	-23	0.000**
	2nd level	3rd level	16	36	-20	0.000**
	2nd level	4th level	16	37	-21	0.000**
	3rd level	4th level	36	37	-1	0.812
Fine	1st level	2nd level	10,000	8,000	2,000	0.387
	1st level	3rd level	10,000	20,000	-10,000	0.003**
	1st level	4th level	10,000	40,000	-30,000	0.000**
	2nd level	3rd level	8,000	20,000	-12,000	0.000**
	2nd level	4th level	8,000	40,000	-32,000	0.000**
	3rd level	4th level	20,000	40,000	-20,000	0.256

*Notes: *p<0.05 **p<0.01*

About the money involved, the figure of sentence length between levels seems to differ less notably, and even the difference between the first level, less than 11970 yuan, and the third level, more than 37000 but less than 115000 yuan, is unremarkable (P=0.078>0.05). This is of course a hidden risk, because the paramount consideration of stealing *PI* is always heading for money, and the higher the amount involved is likely to be associated with the more valuable or richer data, which also constitutes the greater threat to the individual and society. When court judgments are not that differentiated, the effect of warning and punishment will be limited.

Table 6 Dunn's t test pairwise comparison on sentence length and fine for amount of the money involved

	(I) Amount of money involved	(J) Amount of money involved	(I) Median	(J) Median	Difference	*p*
Sentence length	1st level	2nd level	33	19	14	0.213
	1st level	3rd level	33	36	-3	0.078
	1st level	4th level	33	36	-3	0.028*
	2nd level	3rd level	19	36	-17	0.003**
	2nd level	4th level	19	36	-17	0.001**
	3rd level	4th level	36	36	0	0.67
Fine	1st level	2nd level	10,000	17,500	-7500	0.015*
	1st level	3rd level	10,000	40,000	-30,000	0.000**

	(I) Amount of money involved	(J) Amount of money involved	(I) Median	(J) Median	Difference	*p*
	1st level	4th level	10,000	50,000	-40,000	0.000**
	2nd level	3rd level	17,500	40,000	-22,500	0.001**
	2nd level	4th level	17,500	50,000	-32,500	0.000**
	3rd level	4th level	40,000	50,000	-10,000	0.213

Notes: p<0.05 ** p<0.01*

The type of illegal acts is also taken into account. Table 4 shows that there are significant differences in punishments between the acts of only illegally obtaining PI and the other two illegal behavior patterns, and it is a tendency that the legal sanctions for the acts of OIB are less severe, reflecting a judicial philosophy of fully considering the social impact of illegal acts, which is consistent with common sense.

Table 6 Dunn's t test pairwise comparison on sentence length and fine for type of illegal act

	(I) Type of illegal act	(J) Type of illegal act	(I) Median	(J) Median	Difference	*p*
Sentence length	OIO	OSP	16	36	-20	0.005**
	OIO	MSE	16	30	-14	0.000**
	OSP	MSE	36	30	6	0.918
Fine	OIO	OSP	8000	30000	-22000	0.000**
	OIO	MSE	8000	20000	-12000	0.000**
	OSP	MSE	30000	20000	10000	0.085

Notes: p<0.05 ** p<0.01*

5. Discussion

One hundred three judicial documents note that there have been numerous cases of the crime of *IUCPI* via web crawlers in China. This study has concluded the features of subjects of crime, targets of infringement, specific illegal facts, and corresponding judicial judgments through coding and statistical analysis.

At first, it should be stated that Chinese laws have effectively cracked down upon the use of crawlers to infringe upon Chinese citizens' *PI*, and similar cases have been decreasing in recent two years. And based on cases over these years, the study summarized some of the features of persons involved in crimes. Generally, subjects of crime are geographically-dispersed, well-educated, and group-related. The interweaving of the above characteristics is strongly related to the technical attributes of cybercrimes: Firstly, it is the network that allows people to accomplish things together without offline gathering. Therefore, although cases were concentrated in developed eastern regions of China, perpetrators came from all corners of the country. Secondly, not everyone can cross the threshold of programming language and computer technology and develop a crawler effortlessly, and that is why well-educated people with professional knowledge to acquire online *PI* accounted for a higher portion of near 50%. Finally, precisely because it is a type of technology-centric criminal

activity, things like data trafficking require non-technical personnel to do it — criminal gangs formed over time.

As for the target of infringement, with the digitization of Chinese citizens' daily life, information types that can be used to make direct connections with individuals, like phone numbers, have become the main target of spider bots. Interestingly, Chinese lawmakers have always been emphasizing the PI's attribute of identifiability rather than accessibility. Often, criminals collect information only to create connections and profit from it, rather than to figure out who the data of the information is. Understanding the motives and goals of this illegal activity may help adjust legislative and judicial thinking. In addition, the crawler-based *PI* infringement poses a particular challenge to the application of current law in that the amount of the *PI* infringed in such cases easily meets the sentencing standard.

In terms of the legal facts, the whole crime chain of the *IUCPI* via crawlers has been completed, and many defendants almost participated in more than one segment of the crime chain, from tool-providing to data purchasing and selling. Nevertheless, the multi-stage engagement has not been en masse. After all, the technical threshold of programming still exists. In order to prevent the widespread of bad crawlers, China's public security, judicial, legislative, and inspection systems need to work together to manage the diffusion of technology first.

Going back to our core topic, the court outcomes, the research aims to make two points:

First of all, there is no apparent deficiency in the application of current Chinese law. The difference test of court outcomes across various legal facts proves this, and this view is also supported by the decline of cases in recent years. The philosophy of Chinese judicial governance on *PI*, a preventive perspective in nature, implies that, compared with the illegal profit-making on *PI*, China takes the act of *IUCPI* itself into prime consideration - that is, even if 30.90% of the accused have not profited from the *IUCPI* and inflict harm upon society, once a certain number of *PI* has been infringed, Chinese courts will always impose severe sanctions, especially levying fines.

Secondly but most importantly, two common problems of conviction and sentence, the lighter punishment and the insufficient differentiation, cannot be avoided. Lighter punishments are a common occurrence where many accused, or to be more exact, 34 defendants in "especially serious circumstances", have been treated more leniently by courts and been sentenced to less than three years. Among them, 17 people are suspected of violating more than one million pieces of *PI*. Moreover, effective and referential as Chinese legal practices are, China's principle also possibly falls into a legalistic swamp that punishments may not vary much when tens or hundreds of millions of *PI* has been respectively infringed in unrelated cases due to the defined maximum imprisonment. Therefore, web crawlers scraping information speedily and massively, the preventive governance seems inflexible to the inconceivable quantity of stolen *PI*.

6. Conclusion

To sum up, the following aspects are worth considering in future legislative and judicial practices.

Firstly, it is necessary to refresh and update the sentencing standard of *IUCPI*. Crawlers make it easy and fast to obtain personal information, and there already have cases involving hundreds of millions of information. Hence, facing the prevalence of crawler technology, current laws with judicial interpretations may need to adjust the standard of specified amount, no matter for the *PI* or money. If a violation involving 500 million and another involving 50000 personal information are both judged as "especially serious" cases, the former seems to be a type of behavior encouraged:

Now that the crime has been committed, why not take the risk for more benefits?

Secondly, the whole chain of the crime of *IUCPI* via crawlers, along with its separate stage, should be reconsidered. The crime of *IUCPI* comes from the integration of other two crimes, after which the scope of the subject of the crime has been expanded with the statutory punishment aggravated. Apparently, the multi-stage engagement on *IUCPI* comes into being, covering several parts or even the whole chain of crime and relating to other crimes like the telecom fraud with a very real possibility, but did not receive the most severe sanctions. Whether to convict and sentence according to different violation acts may be taken into consideration again.

Finally, though seven defendant units are excluded in this research, the abuse and infringement of citizens' *PI* by companies and enterprises cannot be neglected. On the one hand, the exchange and trade of data among business giants are open facts now, and this type of data transaction may involve the problem of citizens' *PI*. For legislators and judges, big headaches include, but are not limited to, how to judge whether big data deals are illegal or not, and if illegal, how to adjust the existing sentencing standards to apply to the huge amount of data.

Through quantitative analysis, the research maps out the status quo of the *IUCPI* via web crawler in China, and examines convictions and sentences imposed by Chinese courts in prosecutions of invading personal data via crawlers, offering an indigenous empirical window into global legal governance on the relationship between web crawler and *PI*. Furthermore, the research demonstrates the legalistic swamp that the Chinese governance model may fall into, and future research may enlighten us on this matter. We also admit that there still exist some limitations in this study, such as a potential issue of selection bias and missing values of some variables arising from incomplete document recording.

Reference

Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. Journal of economic Literature, 54(2), 442-92.

Baidu Co., Ltd., v. Jietu Co., Ltd., (2016). 262. (CN.)

Brooks, N. (2001). Data Brokers: Background and Industry Overview. Wall Street Journal, 5(5), 552a.

Chellappa, R. K., & Shivendu, S. (2007). An economic model of privacy: A property rights approach to regulatory choices for online personalization. Journal of Management Information Systems, 24(3), 193-225.

China Internet Network Information Center. (2021, September 15). 48th statistical report on China's Internet development. Retrieved October 1, 2021 from http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/.

Criminal Law of the People's Republic of China *2020*. (CN)

Cui, C., & Xu, Z., (2019, June 16). Legal regualtion of web crawlers. Retrieved October 1, 2021 from http://www.cac.gov.cn/2019-06/16/c_1124630015.htm.

Dommett, K. (2019). Data-driven political campaigns in practice: understanding and regulating diverse data-driven campaigns. Internet Policy Review, 8(4).

Epstein, L., & Martin, A. D. (2010). Quantitative approaches to empirical legal research. In Peter Cane & Herbert M. Kritzer (eds.), The Oxford Handbook of Empirical Legal Research. Oxford University Press.

Fahey, R. A., & Hino, A. (2020). COVID-19, digital privacy, and the social limits on data-focused

public health responses. International Journal of Information Management, 55, 102181.
Gao, F., & Wang, W., (2017). The boundary of the crime of selling or providing citizens' personal information-from the perspective of the legal interests protected by the crime of infringing citizens' personal information. Political Science and Law, 02, 46-55. doi:10.15984/j .cnki.1005-9512.2017.02.003.
Gold, Z., & Latonero, M. (2017). Robots Welcome: Ethical and Legal Considerations for Web Crawling and Scraping. Wash. JL Tech. & Arts, 13, 275.
Hong, W., & Jiang Z., (2019). Data Information, Commercialization and protection of personal Information Property Rights. Reform, 03,149-158. doi:CNKI:SUN:REFO.0.2019-03-014.
Imperva. (2020, April 21). Bad Bot Report 2020: Bad Bots Strike Back. Retrieved October 1, 2021 from https://www.imperva.com/blog/bad-bot-report-2020-bad-bots-strike-back/.
Imperva. (2021, April 13). Bad Bot Report 2021: The Pandemic of the Internet. Retrieved October 1, 2021 from https://www.imperva.com/blog/bad-bot-report-2021-the-pandemic-of-the-internet/
Interpretation of the Supreme People's Court and the Supreme People's Procuratorate on Several Issues concerning the Application of Law in the Handling of Criminal Cases of Infringing on Citizens' Personal Information *2017*. (CN.)
Jahankhani, H., Al-Nemrat, A., & Hosseinian-Far, A. (2014). Cybercrime classification and characteristics. In Cyber Crime and Cyber Terrorism Investigator's Handbook (pp. 149-164). Syngress.
Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and Ethics of Web Scraping. Communications of the Association for Information Systems, 47, pp-pp. https://doi.org/10.17705/1CAIS.04724
Li, H., & Sun. H., (2018).On the legal boundary of crawler's behavior of capturing data. Electronic Intellectual Property, (12),58-67. doi:CNKI:SUN:DZZS.0.2018-12-008.
Liu, D., (2007).Property rights protection of personal information.Legal research. 3, 80-91. doi:CNKI:SUN:LAWS.0.2007-03-009.
Liu, Y., (2019). The Research into Criminal Regulation of Web Crawling--from the Perspective of Crimes Against Citizens' Personal Information. Politics and Law, (11), 16-29.
Luo, S., (2020, September 2). The total number of documents in China Judgment Documents Network exceeds 100 million. Retrieved October 1, 2021 from http://www.xinhuanet.com/2020-09/02/c_1126444909.htm.
Lv, B., (2019). Justifying the right to personal information as a civil rights: Taking intellectual property as the reference. China Legal Science, 4,44-65. doi:10.14111/j.cnki.zgfx.2019.04.003.
Marcengo, A., & Rapp, A. (2014). Visualization of human behavior data: the quantified self. In Innovative approaches of data visualization and visual analytics (pp. 236-265). IGI Global.
Murphy, R. S. (1995). Property rights in personal information: An economic defense of privacy. Geo. LJ, 84, 2381.
Ning, L., & Wang, D., (2016).Qualitative analysis of "crawler agreement" of intellectual property rights and its competition law analysis. Jiangxi Social Sciences, (01),161-168. doi:CNKI:SUN:JXSH.0.2016-01-024.
Otto, P. N., Antón, A. I., & Baumer, D. L. (2007). The choicepoint dilemma: How data brokers should handle the privacy of personal information. IEEE Security & Privacy, 5(5), 15-23.

Pangrazio, L., & Sefton-Green, J. (2020). The social utility of ‘data literacy’. Learning, Media and Technology, 45(2), 208-220.

People's Procuratorate of Chen’an County v. Gong, (2018). 179. (CN.)

People's Procuratorate of Haidian District v. Shengpin Co., Ltd., (2017). 2384. (CN.)

People's Procuratorate of of Huai'an v. Niu, (2018). 552. (CN.)

People's Procuratorate of Tiandong County v. Mo, (2018). 230. (CN.)

People's Procuratorate of Yuanjiang City v. Li, (2020), 36. (CN.)

Personal Information Protection Law of the People’s Republic of China’s 2021. (CN)

Phelps, J., Nowak, G., & Ferrell, E. (2000). Privacy concerns and consumer willingness to provide personal information. Journal of public policy & marketing, 19(1), 27-41.

Pittman, P., & Levnberg, K. (2021, June 7). USA: Data protection laws and regulations. Retrieved October 1, 2021 from https://iclg.com/practice-areas/data-protection-laws-and-regulations/usa.

Prins, C. (2006). Property and privacy: European perspectives and the commodification of our identity. Information Law Series, 16, 223-257.

Ruan, L., (2021).The position Standard and Restriction of Criminal Violation of Internet crawler. Hebei Law Science, (07),173-187. doi:10.16494/j.cnki.1002-3933.2021.07.010.

Schwartz, P. M. (2003). Property, privacy, and personal data. Harv. L. Rev., 117, 2056.

Song, X., (2021). Criminal Law Regulation of Abusing Web Crawler Technology to Collect Personal Information. Journal of Hunan University of science and Technology (SOCIAL SCIENCE EDITION)， (04),139-148. doi:10.13582/j.cnki.1672-7835.2021.04.018.

Su, Q., (2021). Evolution of Web Crawling and Conditions for its legitimacy. Comparative Study, (03), 89-104.

Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. Big data, 1(2), 85-99.

Tene, O., Evans, K., Gencarelli, B., Maldoff, G., & Zanfir-Fortuna, G. (2019). GDPR at year one: enter the designers and engineers. IEEE Security & Privacy, 17(6), 7-9.

Varian, H. R. (2002). Economic aspects of personal privacy. In Cyber Policy and Economics in an Internet Age (pp. 127-137). Springer, Boston, MA.

Wang, & Chang. (2020).The Idea Evolution and Thinking Change of Personal Information Protection. Legal system and social development. (06),140-159. doi:CNKI:SUN:SFAS.0.2020-06-009.

Wang, Y., & Chen, K. (2019, August 09). Analysis of typical legal risks and cases of web crawling behavior (Part 1). Retrieved October 1, 2021 from https://mp.weixin.qq.com/s/WF707m8yZTeEXkWZLq4bnA.

Wang, Y., & Chen, K. (2019, August 12). Analysis of typical legal risks and cases of web crawling behavior (Part 2). Retrieved October 1, 2021 from https://mp.weixin.qq.com/s/7IVRGChalqzZjLMaHzzzjQ.

Westin, A. F. (1967). Privacy and freedom. New York: Atheneum.

Whitman, J. Q. (2003). The two western cultures of privacy: Dignity versus liberty. Yale LJ, 113, 1151.

Xia, Y., Zhou, Y., Du, L., & Cai, T. (2020). Mapping trafficking of women in China: Evidence from court sentences. Journal of Contemporary China, 29(122), 238-252.

Xie Y., (2015).The value of personal information from the perspective of information theory with

a review of the privacy protection model. Tsinghua University Law Journal,3,94-110. doi:CNKI:SUN:QHFX.0.2015-03-006.

Xu, G., & Zhang, Z., (2020). The Behavior of Illegal Access to Personal Information of Citizens Should Be Intelligentized, Interpreted and regulated. Journal of people's Public Security University of China (SOCIAL SCIENCE EDITION), (06), 130-142. doi:CNKI:SUN:GADX.0.2020-06-014.

Xu, J., (2020). The Judicial Explanation of Using reptile Technology to Infringe the Legal Benefits of Corporate Data's Intellectual Property Rights. Journal of Suzhou University (PHILOSOPHY AND SOCIAL SCIENCES EDITION), (01), 47-58. doi:10.19563/j.cnki.sdzs.2020.01.007.

Yang, H., (2014). On the influence of crawler protocol on Internet competition. intellectual property right, (01),12-21. doi:CNKI:SUN:ZSCQ.0.2014-01-003.

Yang, Z., (2020).Criminal law regulation of web crawler in the data age. Comparative Study, (04),185-200.

Yeh, C. L. (2018). Pursuing consumer empowerment in the age of big data: A comprehensive regulatory framework for data brokers. Telecommunications Policy, 42(4), 282-292.

You, T., & Ji, L., (2019). Criminal Liability of Possessing Data through Web Crawler-from Shengpin Companys Illegal Acquisition of Computer Information System Data Crime. Application of Law, (10), 3-10. doi:CNKI:SUN:FLSY.0.2019-10-001.

Zhang, P. (2013). The General Provisions and Application of "Anti-Unfair Competition Law"——Thinking about Search Engine Crawler Agreement, Journal of Law Application. 3,46-51. doi:CNKI:SUN:FLSY.0.2013-03-009.

Zhang, T., (2016). Reconstruction of the burden of proof of environmental pollution infringement causality——Based on an empirical analysis of 619 relevant civil judgments. Law Science, 07, 102-111. doi:CNKI:SUN:FXZZ.0.2016 -07-011.

Zhang, Y. (2020, August). Research on Application of Python Web Crawler Technology in DCEP and Legal Risk. In 2020 15th International Conference on Computer Science & Education (ICCSE) (pp. 293-299). IEEE.

Zhou, W., Peng, Y., & Bao, H. (2017). Regular pattern of judicial decision on land acquisition and resettlement: An investigation on Zhejiang’s 901 administrative litigation cases. Habitat International, 63, 79-88.