
# 数据爬取治理报告

上海数据治理与安全产业发展专业委员会
上海赛博网络安全产业创新研究院

2019 年 11 月

本报告系上海数据治理系列沙龙第四期——数据爬取治理的会议成果，由上海数据治理与安全产业发展专委会秘书处单位赛博研究院整理撰写，并得到了专委会诸多专家领导的指导支持。

参与专家：

惠志斌　上海社会科学院互联网研究中心主任

黄道丽　公安部第三研究所网络安全法律研究中心主任

吴卫明　上海市锦天城律师事务所高级合伙人

何　渊　上海交通大学法学院副教授

顾　伟　阿里巴巴法律研究中心副主任

张宝峰　腾讯网络安全与犯罪研究基地高级研究员

朱易翔　翼盾（上海）智能科技有限公司 CEO

撰写团队：

赛博研究院

# 前言

数据爬取作为数据采集的一种高效实现形式，能够在深度和广度上覆盖互联网中大部分的网页链接和内容数据，确保所抽取数据的代表性，极大地降低网络数据检索收集的效率和成本。随着网络空间加速拓展和海量数据蓬勃爆发，更先进的抓取策略、代码架构和技术系统不断被开发应用，数据爬取技术在精准性、时效性、个性化上进展飞速，已经成为国内外诸多互联网企业最为常用的一项技术手段。但是随着围绕数据自动化技术进行爬取和挖掘行为急剧增多，这也带来了数据权属、知识产权和商业机密保护、个人信息隐私界限、不正当竞争、计算机信息系统安全等方面的巨大争议。尤其是近期在金融、内容、电商等各类场景的数据爬取案件频发，对互联网和大数据商业业态产生重大影响，引起了产业界、监管者和全社会的广泛关注。随着我国数据治理法律体系加紧完善和执法力度不断提升，数据爬取这一“灰色地带”也将被提上数据治理议程，其已成为非常紧迫的现实问题。

# 报告核心观点与重要发现

● 数据爬取的核心技术是网络爬虫技术，具备高时效性、高准确性、广覆盖面、低准入门槛和攻防不平衡等特点。基于爬取逻辑可以分为通用式网络爬虫、聚焦式网络爬虫、增量式网络爬虫、深度网络爬虫和分布式网络爬虫。

● 一方面，数据爬取在技术和产业应用上不断发展，极大地促进了数据资源的流通和变现；另一方面，数据爬取的滥用带来了影响网站正常运营，网络安全投入增加、数据泄露风险加剧、新型网络攻击和引发市场恶性竞争等风险。

● 数据爬取是否涉嫌违法犯罪，主要依赖于行为人在数据爬取的“访问进入——获取数据——使用数据”三个阶段下行为的综合考量。主要包括访问进入的计算机信息系统性质，是否得到足够充分授权，是否提供非法程序，访问进入后对计算机信息系统的影响，是否对计算机信息系统安全措施进行技术性规避或破解；是否实质上获取数据，获得数据类型是否涉及个人数据、内容数据、商业数据、重要数据；以及数据据被爬取后的处理方式和流通走向等十一个核心因素。

● 数据爬取在治理中主要面临安全与发展难以平衡，数据基本权属仍无定论，数据法治体系尚不完善，行业性共识规范缺乏，数据壁垒严重，高质量数据供给较少，与其他议题交织复杂等方面的困境难点。

● 数据爬取治理需要在坚持综合治理，完善法制体系，推进数据流通和加强企业合规等方面多个主体共同努力。

# 目录


1 数据爬取概述.....1

1.1 数据爬取的技术原理.....1

1.2 数据爬取的技术特点及分类.....1

1.2.1 数据爬取的技术特点.....1

1.2.2 数据爬取的技术分类.....2

1.3 数据爬取的应用现状及发展趋势.....4

1.3.1 数据爬取的应用现状.....4

1.3.2 数据爬取的发展趋势.....6

1.4 数据爬取滥用的危害.....6

2 数据爬取的现实法律风险.....8

2.1 数据爬取的访问进入.....8

2.1.1 进入.....8

2.1.2 提供.....9

2.1.3 破坏.....10

2.1.4 获取.....11

2.2 数据爬取的对象类型.....13

2.2.1 个人数据.....13

2.2.2 内容数据.....14

2.2.3 商业数据.....15

2.2.4 重要数据.....17

2.3 数据爬取的后续使用.....17

2.3.1 数据交易.....18

2.3.2 数据分析.....18

3 数据爬取治理的困境与难点.....20

3.1 安全与发展平衡需要长期探索.....20

3.2 数据的基本权属问题仍无定论.....20

3.3 完善的数据法治体系尚未建立.....21

3.4 数据产业缺乏行业性共识规范.....21
3.5 数据壁垒加剧数据爬取的乱象.....22
3.6 高质量数据资源供给缺口较大.....23
3.7 数据爬取与其他议题交织复杂.....23
**4 数据爬取治理的原则和建议**.....25
4.1 坚持综合治理原则.....25
4.1.1 平衡发展与监管.....25
4.1.2 加强总体性布局.....25
4.2 加快完善法治体系.....25
4.2.1 完善法治化体系.....25
4.2.2 重视场景化治理.....26
4.3 推进数据有序流通.....26
4.3.1 加强数据高质量供给.....26
4.3.2 消除过度的数据壁垒.....26
4.4 加强企业合规能力.....27
4.4.1 技术提供者.....27
4.4.2 数据爬取者.....27
4.4.3 大数据平台.....27
4.4.5 数据使用者.....28
**5 附录—国内外相关案例**.....29
5.1 百度诉 360 违反爬虫协议案.....29
5.2 大众点评诉百度不正当竞争案.....30
5.3 新浪微博诉脉脉非法抓取用户信息不正当竞争案.....32
5.4 淘宝诉美景大数据产品不正当竞争纠纷案.....33
5.5 央视与视畅公司侵害著作权纠纷、不正当竞争纠纷案.....35
5.6 大众点评网诉爱帮网不正当纠纷案.....37
5.7 CraigslistInc. v. 3Taps Inc., 942 F. Supp. 2d 962.....38
5.8 EarthCam,Inc. v. OxBlueCorp., 703 Fed.Appx.803, 808.....40
5.9 hiQLabs, Inc. v. LinkedIn Corp., 273 F. Supp.3d 1099.....40

# 1 数据爬取概述

## 1.1 数据爬取的技术原理

数据爬取是指通过自动化算法程序，按照一定的规则检索网络空间内容，并从中搜集、提取特定网页数据的过程和行为，其核心技术为网络爬虫（Web Crawler）。

网络爬虫也被称作网络蜘蛛（Web Spider）或网络机器人（Web Robot），其本质是一组脚本或算法程序。首先根据需求目的建立待爬取的URL队列，将精选的种子URL放入队列中，访问其对应的页面并备份数据，同时对页面进行解析并提取所有的其他未被列入队列的URL，将其存入待爬取队列后继续爬取，如此循环往复，直到URL队列中的所有URL爬取完毕或满足系统的一定停止条件为止，在本地或云端上形成的所需数据备份即为网络爬虫的最终结果。

## 1.2 数据爬取的技术特点及分类

### 1.2.1 数据爬取的技术特点

（1）高时效性：数据爬取在获取数据的过程中具备高时效性。相较于人工查找信息，数据爬取通过自动化算法，能够大幅度减少数据查找和获取的时间成本，同时还能够对已获得的数据进行实时更新。在网络空间以TB、甚至未来PB级别的海量数据中，数据爬取将在节省人力、时间和数据更新周期上具备不可比拟的优势。

（2）高精准性：数据爬取在获取数据的结果具备高精准性。由于数据爬取依靠的自动化算法程序规制是可以依照用户目的进行设置的，通过对爬取网页网站和数据检索规则的优化，能够较为精准地

爬取到用户所需要的数据。目前，对于非结构化数据的高精准度爬取，是未来数据爬取技术发展的重点方向。

（3）广覆盖面：数据爬取在获取数据的对象上具备广覆盖面。一般而言，数据爬取的算法规则设置能够涵盖到大部分的公开网站，在数据获取对象的代表性上远远超过人力搜索。

（4）低准入性：数据爬取在技术门槛上具备低准入性。即对于具备计算机编程技能的专业人员来说，利用 Python 语言进行一段简易的数据爬取算法代码编写是相对容易的，逻辑架构和开源代码的公共获取也十分方便。

（5）攻防不平衡：数据爬取在攻防对抗上具备不平衡性。对于反爬取的防守方而言，在技术手段上主要通过访问验证、频率控制、IP 检测等做法，其本质上是通过增加爬虫的成本、降低爬虫的效率来进行对抗，并不能从根本上真正防止数据爬取。而在成本和时间投入上，防守方需要持续性长期投入，远远超过爬取方。

### 1.2.2 数据爬取的技术分类

根据爬取对象的不同，数据爬取可以分为网页爬虫和接口爬虫。网页爬虫是对网络空间的网页超链接进行遍历，来爬取网页的数据信息，也是早期最常见的数据爬虫，最常用于搜索引擎。接口爬虫主要是通过精准构造，形成对特定 API 接口的访问请求，来爬取所需的数据信息，在大数据和移动互联网时代，将会逐渐成为数据爬取的重要方式。

根据爬取逻辑的不同，数据爬取可以分为通用性网络爬虫

(General Purpose Web Crawler)、聚焦式网络爬虫（Focused Web Crawler）、增量式网络爬虫（Incremental Web Crawler）、深层网络爬虫(Deep Web Crawler)、分布式网络爬虫等（Distributed Web Crawler）。

通用性网络爬虫是指通过特定的URL种子延展到整个web网络上进行爬虫的逻辑设计，主要应用在门户站点搜索引擎和大型Web服务提供商采集数据中，爬取范围广，数据量大，但是爬取速度较慢，对存储空间要求高。根据页面爬取优先级，又可细分为深度优先策略、广度优先策略等。

聚焦式网络爬虫是指按照预先定义好的标签规则，有选择性地优先抓取那些与规则相关页面的逻辑设计。其优势在于爬取页面较少，爬取速度和更新频率更快，算力和存储空间等成本消耗较低，需求导向的数据爬取使得结果更具针对性。聚焦爬虫爬行策略实现的关键是对页面内容和链接重要性进行评估，从而圈定特定的爬取优先级和范围，又可细分为基于内容、链接结构评价、增强学习和语境图评价的四种爬取策略。

增量式网络爬虫是指只对新产生的或已经发生变化网页进行数据爬取的逻辑设计。其优势在于能够在规则完善的情况下确保爬取的页面内容是新的（即增量），通常应用于已经抓取了足够数量的页面和数据后，进行数据周期性更新的运营中。它能够有效减少爬取页面和获取数据的重复性，减少时间和资源的消耗。

深层网络爬虫是指对无法通过静态链接进行直接访问的深层网

页内容进行数据爬取的逻辑设计，比如对需要用户注册或提交关键词才能实现访问的数据进行爬取时，就要使用此类爬虫。其优势在于可爬取的网页数量更加庞大，获取数据质量高。

分布式网络爬虫是指同时运行多个逻辑不同的爬虫逻辑，在面对网页规模庞大、结构复杂多样、内容更新频繁时，可以通过对爬虫任务、网页特点、爬取目的等进行分解，进行综合性的数据爬虫。其主要难点在于如何在多个爬虫程序并行的情况下避免重复爬取。

**表 1 基于爬取逻辑的数据爬虫分类**

| 名称 | 逻辑 | 爬取对象 | 特点 |
| --- | --- | --- | --- |
| 通用性网络爬虫 | 最基本常见的爬虫逻辑 | 整个公开 Web 内容 | 覆盖范围广、数据量大，通常应用于搜索引擎和大规模数据采集 |
| 聚焦式网络爬虫 | 按照预先定义好的标签规则，有选择性地优先抓取那些与规则相关页面 | 与规则相关的 Web 内容 | 爬取成本低、页面相关度高、精准性强，通常应用于特定目的的小规模爬取 |
| 增量式网络爬虫 | 对新产生的或已经发生变化网页进行数据爬取 | 未被爬取过的 Web 内容 | 时效性高，通常用于数据更新 |
| 深层网络爬虫 | 对无法通过静态链接进行直接访问的深层网页内容进行数据爬取的逻辑设计 | 无法通过静态链接直接访问的 Web 内容 | 覆盖范围广、爬取数据质量高，通常应用于数据深度分析 |
| 分布式网络爬虫 | 同时运行多个逻辑不同的爬虫程序 | 整个公开 Web 内容 | 复杂程度高、爬取数据质量高，应用于综合性数据需求目的 |

## 1.3 数据爬取的应用现状及发展趋势

### 1.3.1 数据爬取的应用现状

数据爬取最早应用在搜索引擎中，是早期门户网站确保检索内容足够丰富、更新足够及时的核心应用。随着数字经济发展和商业模式不断创新，目前在多个垂直领域都有着广泛实践，包括数字内容、电子商务、互联网金融等各类互联网聚合平台，精准营销、数据分析、

风险控制等各项业务，具体见表 1。

**表 2 互联网经济中数据爬取的典型应用**

| 行业 | 业务 | 具体应用 |
| --- | --- | --- |
| 金融 | 风险控制 | 基于云平台、云计算、机器学习等技术，通过数据采集能力和数据挖掘能力，对用户进行风险评估，帮助企业客户深度剖析其用户的诚信画像。 |
| 电子商务 | 精准营销 | 通过采集用户消费数据，分析用户消费偏好，并据此向用户推送有针对性的商品信息。 |
| 金融 | 信贷催收 | 提供精准营销模型、反欺诈、多维度用户画像、授信评分、贷后预警、催收智能运筹等全面风险管理。 |
| 票务 | 机票抢订 | 通过爬虫技术抢占低价票，随后寻找真正客源，将低价票加价转售。 |
| 公共舆情 | 舆情监测分析 | 通过爬取论坛、微博、微信内容，自动分析出色情、暴力等信息提交。 |
| 公共管理 | 税务稽查 | 按照税务稽查所下达查验指令，快速抓取税务稽查人员索要的分析结果和反映出纳税人涉税异常点，有效查处部分偷逃税现象。 |
| 财会行业 | 审计 | 模拟浏览器向网络服务器发送请求，将网络资源从网络流中读取出来，再对这些信息进行简单提取，从中得出问题线索。 |

Distil Networks 发布的《2019 年不良机器人报告》（Bad Bot Report 2019）显示，2018 年全球流量中，真人流量占比 62.1%，机器人流量总共占比 37.9%，其中坏机器人流量占比 20.4%，好机器人流量占比 17.5%。而高级持久性机器人在整个机器流量占比 73.6%，自动化数据爬取技术对于人类访问模式的模仿程度日益逼真，这使得全球真实的机器流量占比可能更高。而在垂直行业，金融（42.2%）、

票务（39.3%）、教育（37.9%）、IT 与服务（34.4%）和市场营销与广告（33.3%）分别是坏机器人行业流量前五位。

这意味着数据爬取已经成为了部分行业中数据获取和流通的主要方式，未来可能会成为互联网时代几乎所有平台和企业获取量级网络数据的基础手段。

### 1.3.2 数据爬取的发展趋势

（1）技术上，数据爬取呈现难识别、低门槛、个性化的发展模式。仿真度不断提高，识别对抗困难，爬虫技术门槛进一步降低，个人更容易获得各种低量级、低频度的爬取工具和爬虫代码，爬虫与自动访问界限模糊，针对行业级 APP 和数据库定向开发的数据爬取不断增多。

（2）应用上，数据爬取的应用规模在不断扩大。目前，以金融行业的风险控制、精准贷款，电子商务的票务购买、数字营销，数字内容的内容推送、内容抓取为主要场景。同时，数据爬取滥用和恶用带来的垃圾流量增长更快，影响到互联网经济正常发展。

（3）监管上，数据爬取的法律风险在不断加大。我国数据安全法律密集出台，针对互联网金融信贷领域的数据安全问题频出重拳，但是数据爬取的法律边界和治理体系亟需明确完善，企业使用数据爬取进行商业行为的法律风险在不断上升。

## 1.4 数据爬取滥用的危害

数据爬取作为互联网大数据时代应运而生的自动化数据获取方式，在促进网络空间数据高效流通的同时，也带来了相应的风险危害，

主要表现为技术性直接危害和社会性外溢风险。

**技术性直接危害是指数据爬取会直接影响被爬取网站的正常运行，增加运营成本。**自动化数据爬取意味着短时间内的大规模站点访问和数据缓存，将会直接在算力、流量传输、数据存储等方面加大网站的运行压力，由此而产生的“垃圾流量”可能会使得互联网运营者产生错误决策，甚至挤压用户正常访问空间，降低用户体验甚至导致无法正常使用；另一方面，数据爬取会直接提高网络运营者的运营成本和安全成本，运营者在数据爬取情况下要保证网站的正常运行，必然要加大运营的管理和资金投入，或者利用反爬虫技术进行对抗，提高安全投入。

社会性外溢风险包括三个方面，**一是加大数据安全风险。**数据爬取的根本目的在于获取数据进行再生产或流通，使用数据爬取从网站收集数据后，将会导致数据脱离原有网站，成为数据泄露的源头，从而引发各类安全和法律问题。**二是可能成为新型网络攻击方式。**恶意使用者可能滥用数据爬取，将其作为一种对网站服务器的站点攻击方式，使得一些部署在云端或者服务器较小的网站产生卡顿、瘫痪。**三是造成市场恶性竞争。**数据是互联网企业的核心资产和竞争力，数据爬取的泛滥和反制对抗将会阻碍数据正常流通，冲击市场秩序，恶化互联网经济业态，造成商业资产、机密流失或不正当竞争。

## 2 数据爬取的现实法律风险

数据爬取在场景中的实践中可大致分为“访问进入——获取数据——使用数据”三个阶段，数据爬取是否涉嫌违法犯罪主要依赖于行为人在三个环节下行为的综合考量。

其中在“访问进入”阶段，主要影响因素包括访问进入的计算机信息系统性质、是否得到足够充分授权、是否提供非法爬虫程序、访问进入后对计算机信息系统的影响、是否对计算机信息系统安全措施进行技术性规避或破解、是否实质上获取数据六个个方面。在“获取数据”阶段，主要影响因素是获取数据对象的类型、属性和规模，在个人数据、内容数据、商业数据、重要数据四种数据类型下，可能会产生违法犯罪后果。在“使用数据”阶段，主要影响因素是指数据被爬取后的处理方式和流通走向，在不同行业不同业务下需要具体考虑，对于现行法律规制违法犯罪行为提供数据爬取支持，可能会面临连带责任或数罪并罚。

### 2.1 数据爬取的访问进入

#### 2.1.1 进入

使用数据爬虫技术进入到计算机信息系统内部行为本身就可能构成违法犯罪，**核心要素在于被进入的计算机信息系统性质以及是否具备足够充分授权。**比如涉及到国家安全和国家秘密的政府内网、国防建设、尖端科学技术领域的计算机信息系统，只要实施了非法侵入行为即构成非法侵入计算机信息系统罪。

在无授权情况或超出授权范围使用账号、密码登录计算机信息系

统，也可能被视为侵入计算机信息系统的行为，构成犯罪行为。访问授权在各类内网系统比较明确，但是对于网络空间中公开存在网站的访问权限界定却比较模糊，网络运营者的 Robots 协议、事先拒绝爬虫访问声明并不具备事实上的法律效力，却可能成为司法判定中的重要参考。

【相关法律】《刑法》第二百八十五条 非法侵入计算机信息系统罪。

【案例】：滕守昆、蒋小东、房川、滕守灿四人非法侵入计算机信息系统案

该案中，被告人滕某等四人为了在帮人处理车辆交通违章业务时为方便查询相关车辆的信息，违规在其手机上下载“四川公安交警警务云平台”APP 软件并进行安装，再通过非法获取的相关用户名及密码登录网站平台，对相关车辆的交通违章等信息进行非法查询，非法查询相关信息一千余条。经过四川省公安厅网络安全保卫总队认定，《四川公安交警警务云平台》和四川公安交警警务云平台 APP 计算机信息系统属于国家事务类网站，滕某等四人于 2017 年 11 月被攀枝花市仁和区人民法院认定为构成非法侵入计算机信息系统罪。

### 2.1.2 提供

供应商和技术人员提供数据爬取技术的行为也有可能构成违法犯罪，**核心要素在于是否提供了专门用于侵入计算机信息系统的程序，或者明知他人实施侵入计算机信息系统的违法犯罪行为而为其提供程序。**根据 2011 年 6 月“两高”《关于办理危害计算机信息系统

安全刑事案件应用法律若干问题的解释》的相关规定，提供具备下列技术特征的，构成提供侵入、非法控制计算机信息系统程序、工具罪。一是具有避开或者突破计算机信息系统安全保护措施，未经授权或者超越授权获取计算机信息系统数据的功能的；二是具有避开或者突破计算机信息系统安全保护措施未经授权或者超越授权对计算机信息系统实施控制的功能的；三是其他专门设计用于侵入、非法控制计算机信息系统、非法获取计算机信息系统数据的程序工具。

【相关法律】《刑法》第二百八十五条 提供侵入、非法控制计算机信息系统程序、工具罪。

【案例】马小辉、王贵兴侵犯公民个人信息案

该案中，被告人张大云根据马某的用户需求，伙同他人开发制作“林某1”、“凌某”系列软件并出售给马某，供其在未获得淘宝公司授权、账号权利人许可的情况下，故意规避淘宝、支付宝公司采用的安全防护系统批量获取用户个人数据信息，并在知情情况下继续进行软件的开发和维护行为，并从中获利，被浙江省绍兴市中级人民法院2017年3月依法判处提供侵入、非法控制计算机信息系统程序、工具罪。

### 2.1.3 破坏

使用网络爬虫技术可能会对被爬取的网站或数据造成外溢性影响，情节严重的可能会构成违法犯罪。**核心要素在于访问进入后对计算机信息系统的影响和是否对计算机信息系统安全措施进行技术性**

**规避或破解。**比如进行爬虫过程中对计算机信息系统造成了较为严重的数据破坏或程序干扰，或者存在对计算机信息系统安全技术措施的暴力破解和规避绕取，甚至将爬虫技术滥用为网络攻击方式，对系统中存储、处理或传输的数据进行删除、篡改等行为，都将可能会构成破坏计算机信息系统罪。

【相关法律】《刑法》第二百八十六条 破坏计算机信息系统罪。

《数据安全管理办法（意见征求稿）》第十六条 网络运营者采取自动化手段访问收集网站数据，不得妨碍网站正常运行；此类行为严重影响网站运行，如自动化访问收集流量超过网站日均流量三分之一，网站要求停止自动化访问收集时，应当停止。

【案例】王博一文、黄业兴破坏计算机信息系统案

在该案中，被告人使用电脑通过 Python 软件编写“爬虫”程序，并将其编写的“爬虫”程序植入全运会组委会接待服务管理系统进行攻击，删除了该系统内大量参赛运动员及技术官员的抵离信息、酒店住宿信息、人员简要身份信息，致使当日天津市全运会组委会接待服务部 39 台计算机无法正常运行接待服务系统，被天津市第一中级人民法院依法判定为破坏计算机信息系统罪。

### 2.1.4 获取

使用爬虫数据非法进入到计算机信息系统，并获取了该计算机信息系统中存储、处理或者传输的数据，或者对该计算机信息系统实施非法控制，可能会构成违法犯罪。**核心要素在于是否在事实上获取数**

**据。**通常而言，被认定为非法侵入且获取数据后，同视为非法获取计算机信息系统数据罪。而在合法进入或合法性不明情况下进入计算机信息系统时，是否构成非法获取计算机信息系统数据罪取决于进入方式和获取数据类型。

【相关法律】《刑法》第二百八十六条 非法获取计算机信息系统数据、非法控制计算机信息系统罪。

《网络安全法》第二十七条 任何个人和组织不得从事非法侵入他人网络、干扰他人网络正常功能、窃取网络数据等危害网络安全的活动；不得提供专门用于从事侵入网络、干扰网络正常功能及防护措施、窃取网络数据等危害网络安全活动的程序、工具；明知他人从事危害网络安全的活动的，不得为其提供技术支持、广告推广、支付结算等帮助。

【案例】上海晟品网络科技有限公司、侯明强等非法获取计算机信息系统数据罪案

在该案中，被告人侯某作为被告单位的技术主管人员（CTO），指示被告人郭某采用技术手段破解北京字节跳动网络技术有限公司服务器的加密算法和API交互规则，并使用伪造的设备ID绕过服务器的身份校验和访问频率设置等规避手段，非法获取被害单位服务器当中存储的视频数据并造成被害单位损失技术服务费2万元，被北京市海淀区人民法院2017年11月依法判处非法获取计算机信息系统数据罪。

## 2.2 数据爬取的对象类型

### 2.2.1 个人数据

根据两高《关于办理侵犯公民个人信息刑事案件适用法律若干问题的解释》，个人信息是指以电子或者其他方式记录的能够单独或者与其他信息结合识别自然人个人身份的各种信息，包括但不限于自然人的姓名、出生日期、身份证件号码、个人生物识别信息、住址、电话号码等。其中，可识别性是个人信息判定的核心要素。在未经被数据所有者同意的情况下爬取用户个人数据的行为，不管是个人隐私数据还是一般数据，均可能会构成侵权或犯罪。在实践中通常会有两种情况，一是公民在网络空间中主动公开发布、暴露的个人数据，其在司法判定中并不必然构成被爬取的充分授权依据，爬取公开状态的个人数据依旧具有法律风险。二是对于公民未公开个人信息的爬取，将大概率构成侵犯公民个人信息罪。目前，对于个人数据的违法爬取和利用是数据爬取治理的核心重点，对于个人数据的采集、使用不得侵犯公民个人权利，不得使公民个人信息处于危险状态。

【相关法律】《民法总则》第一百一十一条 自然人的个人信息受法律保护。任何组织和个人需要获取他人个人信息的，应当依法取得并确保信息安全，不得非法收集、使用、加工、传输他人个人信息，不得非法买卖、提供或者公开他人个人信息。

《刑法》第二百五十三条 侵犯公民个人信息罪。

《网络安全法》第四十一条 网络运营者收集、使用个人信息，应当遵循合法、正当、必要的原则，公开收集、使用规则，明示收集、使

用信息的目的、方式和范围，并经被收集者同意。网络运营者不得收集与其提供的服务无关的个人信息，不得违反法律、行政法规的规定和双方的约定收集、使用个人信息，并应当依照法律、行政法规的规定和与用户的约定，处理其保存的个人信息。

【案例】于剑、宁某侵犯公民个人信息案

在该案中，被告人于剑等人作为“同信缘”小额贷款平台的合伙人为牟利，合谋开发某“黑爬虫”网站。该网站利用爬虫技术非法获取多家小额贷款平台内公民个人借贷信息、身份证照片信息等公民个人信息，用户充值后，经付费可以通过“黑爬虫”网站查询公民个人信息，并为付费查询“同信缘”等数家小额贷款平台提供公民个人借贷信息以及公民身份照片查询功能和数据采集接口。2017 年 9 月至 2018 年 7 月，相关网站的公民个人借贷信息、身份证照片信息累计查询量已达 84 万余次。被告人于某、宁谋被依法判定侵犯公民个人信息罪。

### 2.2.2 内容数据

内容数据通常指发布者通过劳动搜集、整理、汇编、创作，具备稀缺性、独创性和商业价值的作品，以文字、图片、音频、视频、用户评论、数据库等各种形式呈现。在未经被数据所有者同意或授权的情况下，通过网络爬虫技术爬取数字内容网站平台上的各类内容数据，均有可能构成著作权侵权甚至侵犯著作权罪。比如，使用网络爬虫对他人信息进行爬取，可能侵犯内容所有者的“复制权”，爬取后进行公开传播可能会侵犯内容所有者的“信息网络传播权”，爬取后

进行商业营利行为可能会构成侵犯著作权罪。但在实践中，侵权行为的判定通常要结合内容所有者的主观意图、平台协议、爬取后的内容呈现方式等因素综合考量。

【相关法律】《著作权法》第十条 （十二）信息网络传播权，即以有线或者无线方式向公众提供作品，使公众可以在其个人选定的时间和地点获得作品的权利。

《刑法》第二百一十七条 侵犯著作权罪，是指以营利为目的，未经著作权人许可复制发行其文字、音像、计算机软件等作品，出版他人享有独占出版权的图书，未经制作者许可复制发行其制作的音像制品，制作、展览假冒他人署名的美术作品，违法所得数额较大或者有其他严重情节的行为

【案例】深圳市腾讯计算机系统有限公司与北京字节跳动科技有限公司侵害作品信息网络传播权纠纷

在此次纠纷中，法院认定，字节跳动公司未经腾讯公司许可，在其经营的今日头条网上使用了腾讯公司依法享有著作权的文字作品共 4783 字内容（《高峰：在看守所的六个月是财富》），使公众可以在其个人选定的时间和地点获得涉案作品部分内容，侵害了腾讯公司享有的信息网络传播权，字节跳动公司应当对其侵权行为承担相应的法律责任。

### 2.2.3 商业数据

商业数据包括商业秘密和商业数据资产，前者是指不为公众所知

悉、能为权利人带来经济利益、具有实用性并经权利人采取保密措施的技术信息和经营信息；后者则在涵盖前者的基础上，包括了在公开网络环境存在，能够给权利人带来市场竞争力和经济利益、具备资产投入和维护的数据信息。在实践中，商业数据被视为企业资产，因此对于商业数据的爬取可能会构成侵犯商业机密违法犯罪行为或不正当竞争，具体判定因素包括爬取行为人与被爬取人的市场关系、爬取的商业数据是否具备秘密或资产属性等。

【相关法律】《刑法》第二百一十九条 侵犯商业秘密罪，是指以盗窃、利诱、胁迫或者其他不正当手段获取权利人的商业秘密，或者非法披露、使用或者允许他人使用其所掌握的或获取的商业秘密，给商业秘密的权利人造成重大损失的行为。

《反不正当竞争法》第九条　经营者不得实施下列侵犯商业秘密的行为：(一)以盗窃、贿赂、欺诈、胁迫、电子侵入或者其他不正当手段获取权利人的商业秘密。

【案例】深圳市谷米科技有限公司与武汉元光科技有限公司等不正当竞争纠纷案

在该案中，被告元光公司存在利用网络爬虫技术大量获取并且无偿使用原告谷米公司“酷米客”软件的实时公交信息数据的行为。公交车作为公共交通工具，其实时运行路线、运行时间等信息仅系客观事实，但当此类信息经过人工收集、分析、编辑、整合并配合 GPS 精确定位，作为公交信息查询软件的后台数据后，其凭借预报的准确度和精确性就可以使“酷米客”APP 软件相较于其他提供实时公交信息

查询服务同类软件取得竞争上的优势。因此，谷米公司的实时公交数据被认为具有实用性并能够为权利人带来现实或潜在、当下或将来的经济利益，其已经具备无形商业资产的属性。因此，被告的数据爬取行为被法院认定为具有非法占用他人无形财产权益，破坏他人市场竞争优势，并为自己谋取竞争优势的主观故意，违反了诚实信用原则，扰乱了竞争秩序，构成不正当竞争行为。

### 2.2.4 重要数据

重要数据是指不涉及国家秘密，但如果泄露、窃取、篡改、毁损、丢失和非法使用可能危害国家安全、国计民生、公共利益的未公开数据，包括且不限于地理、自然资源、重要物资储备等数据；基因、生物特征、疾病等数据；宏观统计等重要经济数据；网络信息系统的缺陷、漏洞、防范措施等数据；人群导航位置、大型设备目标位置和移动数据；法律法规规定的其他重要数据。对于重要数据的广泛爬取可能会直接影响国家安全、经济安全、社会稳定、公共健康和安全，因此，《数据安全管理办法（意见征求稿）》第十五条规定，网络运营者以经营为目的收集重要数据或个人敏感信息的，应向所在地网信部门备案。目前，对于重要数据的认定和相关法律尚未完善，但是随着未来我国政府公共数据的大规模共享开放，对于重要数据面临自动化爬取风险的治理将会提上日程。

## 2.3 数据爬取的后续使用

大数据时代，数据的高速流转和运营创造了空前的商业价值，数据爬取则是主要数据来源之一。目前在实践中，国家重点打击和高风

险的数据使用方式主要包括：为数据黑产和非法交易提供原始数据来源的数据爬取行为；在金融领域为风险控制、精准放贷和线下催债提供支撑的数据爬取行为；在数字内容领域基于商业目的进行内容公开传播的数据爬取行为；基于损害所有者市场竞争力、产生替代性或竞争性市场活动的数据爬取行为等。

### 2.3.1 数据交易

近年来数据倒卖猖獗，数据非法交易案件频繁，并呈现产业链作案特征。数据爬取遂成为数据黑产的上游，为各类数据中间商、代理商和最终使用者提供数据来源。同时，数据交易不但能够使得数据商业价值变现，也是数据“洗白”的关键流程，通过数据清洗、拖库撞库、数据整合等技术手段，各类非法获取的数据实现了在公开渠道上的合法流通。目前，我国尚未有专项规制数据交易的法律体系，数据交易的法律风险主要在于数据来源方式、数据类型等。如果被传播的数据内容系淫秽物品，或用于牟利的数据内容系淫秽物品，或者是对公民个人信息实施的出售、非法提供行为，或者使用或允许他人使用其所掌握的或获取的商业秘密，达到所规定行为后果的，则会存在构成传播淫秽物品罪、传播污秽物品牟利罪、出售、非法提供公民个人信息罪以及侵犯商业秘密罪等法律风险。如果是知识产权保护范围内的数据，侵害著作权罪则是最主要的刑事风险。

### 2.3.2 数据分析

在互联网金融借贷行业中，各类风控企业往往依靠数据爬取获得大量用户数据，从而对用户进行精准画像进行风险评估，不但侵犯公

民隐私，甚至借助用户数据中的地理信息、联系人信息等进行线下暴力催收，可能会构成违法犯罪。近期同盾科技、魔蝎科技、新颜科技、公信宝、聚信立、白骑士等诸多知名大数据金融风控公司的数据爬虫业务相继主动或被动关闭，其中主要涉及的违法情况就是大数据支持的高利贷和线下暴力催收等违法行为。

## 3 数据爬取治理的困境与难点

在大数据时代，数据的充分流通和价值变现是数字经济发展的核心，数据爬取将是各类市场主体获取数据不可替代的自动化工具，数据爬取将面临更加复杂和现实的治理难点。

### 3.1 安全与发展平衡需要长期探索

数据爬取问题本质是数字经济发展中带来治理难题，数据爬取治理要在保障数字经济高质量发展和数据安全保障之间取得平衡。在数据海量爆发的时代，对于数据获取的精准性、高效率、广覆盖、时效性的要求越来越高，数据爬取将是各类市场主体获取数据不可替代的自动化工具，因此对于数据爬取的监管不能“一刀切”。另一方面，越来越多的国家、社会和个人信息以数据形式暴露在网络空间，数据安全问题在现代国家治理重要性日益突出，数据爬取的监管规制将成为数据治理的重要一环。目前，数字经济仍处于产业高速发展、形态迅速变化、技术不断创新的早期阶段，如何在安全与发展之间寻求平衡点，需要各方长期的探索。

### 3.2 数据的基本权属问题仍无定论

数据归属和权利问题未定是影响数据爬取治理的主要瓶颈。具体是指包括数据的所有权、使用权和收益权等在内的数据各项权利归属哪类主体，并延伸至数据权利确认、数据主体监管、数据滥用禁制等数据权利法律原则怎样建立等一系列问题。当数据权属不定时，获得数据所有者的足够充分授权后进行数据爬取将无从实现，或将直接造成法律纠纷。比如，由于公民在网络空间中主动公开发布、暴露的个

人数据，其在司法判定中并不必然构成被爬取的充分授权依据，企业是否能够直接爬取公民主动发布在网络空间的个人公开数据，以及如果需要足够充分授权应当由谁来授权等此类问题，已经成为实务中相关案件爆发的主要诱因点。如在微博诉今日头条一案中，微博认为用户在其平台发布的内容数据属于微博，在未授权情况下被今日头条非法爬虫，而今日头条主张此类内容数据属于用户而非微博，在得到用户充分授权后，对用户在微博发布的内容进行爬虫不具有违法性。

### 3.3 完善的数据法治体系尚未建立

目前，我国尚未有规制数据爬取的专项法律或部门规章，数据安全相关法律法规依然在筹备和意见征求阶段。对于数据爬取的治理主要依靠国家网信部门和公安机关通过专项执法活动直接打击数据爬取的相关违法犯罪行为，和司法机关通过《民法》《刑法》《网络安全法》《反不正当竞争法》《著作权法》等相关条款进行事后兜底治理。完善的数据法治体系尚未建立使得数据爬取存在巨大的“灰色地带”。市场主体数据爬取行为的法律边界不清晰，“运动式”执法监管模式难以形成真正持续、良性、有序的治理范式，使得数据爬取治理缺乏统筹性的制度建设。

### 3.4 数据产业缺乏行业性共识规范

数据爬取主要以市场主体的经济行为为主，通用性共识和行业性规范形成困难也是治理难点之一，表现为已有的通用性规范的法律认可度低，新的行业性规范难以形成。以目前业界通行的 Robots 协议为例，Robots 协议是技术界为了解决爬取方和被爬取方之间通过计

算机程序完成关于爬取的意愿沟通而产生的一种机制，由网站所有者通过位于置于网站根目录下的文本文件 Robots 提示网络机器人哪些网页不应被抓取，哪些网页可以抓取。按照现行惯例，未被 Robots 协议排除的数据属于互联网上的公开数据，任何人都有权访问和收集。如果违反 Robots 协议，强行爬取他人的数据，则可能被认定为违反诚实信用和商业道德，构成不正当竞争。但是，在 Robots 协议已获得中国互联网协会《互联网搜索引擎服务自律公约》背书情况下，其在执法实践和司法判定中的法律效力依旧争议颇多，法律效力也未被正式认定，可行度低。同时，在缺乏有效通行的行业规范情况下，互联网企业纷纷倾向于通过事前声明或用户协议直接拒绝数据爬取，行业内对话和谅解困难，共识规范难以达成。

### 3.5 数据壁垒加剧数据爬取的乱象

在现实不少案件中，经营者之所以采取爬虫技术抓取数据，往往是因为面临对数据的强烈需求与难以从正常商业渠道获取数据的矛盾情形。比如在深圳“酷米客”诉“车来了”案件中，基于 GPS 定位设施提供商与公交公司之间签订的《GPS 设备安装协议》，并且明确拒绝后进入者的数据交易请求，实际上导致其他经营者无法获取相应数据。在行业内数据资源分布极不均衡，数据大量集聚在垄断性数据平台和公司已经成为事实的情况下，一方面“数据垄断”产生了在法律、渠道、技术上多层次的“数据壁垒”，尤其是部分行业特定的、具备稀缺性且不可替代的数据过度集中在互联网巨头中，使得数据垄断成为事实上的商业垄断，成为数据爬虫乱象加剧的深层经济根源。

另一方面，不直接生成数据或者生成数据十分有限的企业有着较为强烈采取数据爬取手段挖掘数据的内因驱动。在市场规律作用下可能会产生一系列的负外部性，如不劳而获地“搭便车”行为，进一步恶化市场竞争秩序。对当下游走于合法与非法的灰色地带使用技术手段突破数据获取壁垒的企业而言，更为棘手的问题是，移动互联网的发展使得主要数据获取方式从相对开放的浏览器转向更为封闭的 APP，导致数据爬取还面临各 APP 反爬加密技术的技术壁垒限制，从而导致“灰色”爬取越来越可能涉嫌非法获取计算机信息系统数据的违法犯罪。

### 3.6 高质量数据资源供给缺口较大

我国良性、有序的数据流通体系总体处于起步阶段，政府部门的公开数据开放不足，合法交易平台和机制尚未大规模普及推广，合法公开渠道的数据供给在数量和质量上都无法满足数字经济快速发展的需求，导致部分企业“被迫”采用非法爬虫技术获取数据。其背后深层次原因，仍在于数据权属未定和现实的数据壁垒。这使得可能存在法律风险的各类数据因其权属问题未定而暂时无法进入到合法交易渠道中，而互联网企业事实上倾向于形成内部封闭的“数据孤岛”，部分中小企业被迫采用非法爬虫技术获取数据。

### 3.7 数据爬取与其他议题交织复杂

数据爬取与信息时代的其他问题交织，使得治理复杂程度和难度不断攀升。首先，数据爬取可能会造成数据的跨境流动，使得在治理实践中面临更多的政策障碍，比如以阿里、腾讯为代表的中国互联网

企业在开拓海外业务时面临的异地数据爬取问题，和以苹果、谷歌等为代表的西方互联网企业在我国境内采集数据的问题等。其次，数据爬取与以大数据、人工智能等为代表的智能化技术结合，使得数据爬取更加难以识别，技术性安全措施难以对抗，而数据爬取后续的数据清洗、变现途径更加多样，维权取证困难。最后，以物联网、5G 等为代表的新一代信息基础设施升级改造带来更庞杂、异构、海量的数据来源，使得数据爬取未来可能与物联网安全深度交织，安全治理难度加大。

# 4 数据爬取治理的原则和建议

## 4.1 坚持综合治理原则

### 4.1.1 平衡发展与监管

数据爬取治理是大数据经济在发展过程中产生的治理问题，涉及到公民权益、网络空间安全和数字经济发展等各个方面，要充分发挥网信部门统筹协调功能，加强与产业部门、公安部门和司法机关的协调配合，平衡好数据保护监管和数据产业发展之间的关系。既要发挥数据爬取在提高数据获取效率、推动数据流通共享的积极作用，也要规制数据爬取滥用带来一系列安全、社会问题，在发展与监管之间，探索出平衡的治理路径。

### 4.1.2 加强总体性布局

数据爬取是数据价值实现必不可少的基础方式，与数据聚合、流通和变现密切相关。因此，要在整个数据生命周期总体性布局下加强数据爬取的针对性治理，包括加强对数据基本权属问题的研究探索，加强数据爬取与其他行业发展、网络安全等交织问题的前瞻性布局，在数据爬取的技术不断发展、产业不断演变中形成长效性治理机制，在发展的过程中，解决治理问题。

## 4.2 加快完善法治体系

### 4.2.1 完善法治化体系

要加快完善数据治理法治体系，推动数据爬取的专项立法。综合运用执法和司法手段，对于侵害公民权益和危害国家安全的违法犯罪行为予以坚决打击。同时，要坚持刑法的“谦抑性”原则，避免过重、

过度入刑，特别是对移动互联网时代越来越普遍的 APP 逆向破解，需要结合技术发展趋势，审慎判断是否涉嫌违法犯罪。同时，结合数据爬取实践，加强网络安全法、著作权法、反不正当竞争法等解释性应用。

#### 4.2.2 重视场景化治理

基于垂直领域下的“场景化治理”逐渐成为数据治理较为有效的方式。因此，对于数据爬取的正当性、违法性判定，要在具体场景、具体行业下针对具体行为综合考量。在数据爬取广泛应用的互联网金融、电子票务、数字内容、电子商务、精准营销等业务场景下率先形成行业性数据爬取规范标准，支持各利益相关方积极探索数据爬取通用性共识。

### 4.3 推进数据有序流通

#### 4.3.1 加强数据高质量供给

政府要带头推进数据开放，以开放为原则、不开放为例外，向市场主体高质量地开放各类数据资源。要大力推进数据良性流通机制建设，推广数据合法交易平台体系普及推广，提高政府部门公共数据的有序开放，加快对于各类数据资源的确权、定价，压缩数据灰产、黑产的经济空间，引导更高质量、更广范围的数据进入到合法供给渠道，满足数字经济发展需求。

#### 4.3.2 消除过度的数据壁垒

数据作为大数据时代的核心资源和资产，过度的行业壁垒、技术壁垒形成的“数据孤岛”现象将阻碍数据价值在更深层的聚集实现。

采取有效激励措施，鼓励平台企业和数据所有者积极主动开放能够促进中小企业发展的高质量商业数据和必要的行业准入数据，对于那些具有稀缺性的不可替代数据，探索建立特定数据所有人或持有人开放行业必要数据的机制。

## 4.4 加强企业合规能力

### 4.4.1 技术提供者

基于爬虫技术与反爬虫技术的不平衡性，在审慎适用“技术中立性”的避风港原则的同时，要求技术提供者对数据爬取工具的开发在访问频度、爬取对象等参数上进行必要的技术限制，对爬取工具的采购方进行必要审核和提示，发现滥用爬取工具甚至用于违法犯罪活动的，应当依法停止提供并向有关部门举报。

### 4.4.2 数据爬取者

对于数据爬取者而言，要坚持合法访问第三方计算机信息系统，遵守 Robots 协议，获得权利人足够充分的授权，降低数据爬取对于被爬方的外溢影响，在法律边界内获取相关的数据类型。同时，对于数据后续的流向和使用，要加强审核监督，避免合理获取的数据被用于不合理用途。

### 4.4.3 大数据平台

要充分保护好用户在平台网站的数据安全，加强反爬技术的部署深度，明晰平台对于数据是否可被爬取、怎样合法爬取、可爬取的时段和规模的态度，做到事前声明。同时，从促进产业生态发展角度出发，积极主动通过开放平台、应用市场等开放数据，促进产业生态的

繁荣发展。

#### 4.4.5 数据使用者

要坚持使用合法数据来源建立对于数据来源的技术和法律审查流程机制，避免误用非法爬取的数据资源。坚决抵制数据来源不清或者非法手段获取的数据。同时，要依法建立合理利用数据的内部制度与流程体系，确保数据处理全生命周期的合法合规。

# 5 附录—国内外相关案例

## 5.1 百度诉360违反爬虫协议案

### 5.1.1 主要事实

2012年8月，奇虎公司（360）推出新的搜索引擎，并通过360搜索和浏览器强行抓取百度等搜索引擎内容。百度公司认为奇虎公司违反了百度网站上设定的爬虫协议，非法爬取了百度网页内容，抄袭了百度百科词条，以及恶意拦截篡改百度旗下运营的网站，同时安插后门泄露用户数据隐私。百度还指责奇虎公司将用户的自主搜索引擎替换成另一家搜索，成为网址导航的默认搜索引擎。但360认为，百度受幕后黑手指使采取抹黑行为，在乌云漏洞平台上发布虚假漏洞公告。在中国互联网协会的组织下，2012年11月1日，多家互联网企业在北京发起签署了《互联网搜索引擎服务自律公约》，强调互联网服务提供者需遵守爬虫协定。签订当日，百度向奇虎公司发送了法律通知函，要求奇虎公司停止抓取、复制百度网站的页面和数据内容。

### 5.1.2 裁判要点

360违反爬虫协议的行为，是否造成了不正当竞争。法院认为，互联网经营者遵循自愿平等公平诚信的原则，互联网发展依赖自由竞争，但归根结底在自由竞争下实现创新，在法律准许范围内竞争得到维护，但并非毫无规范的竞争，采用违反商业道德的行为市场会停留在丛林法则的误区，要强调规则的重要性。爬虫协议应当被认定为搜索引擎行业内公认的、应当被遵守的商业道德。360违反爬虫协议，擅自抓取百度网站内容并生成快照向用户提供，已明显超出网页快照

的合理范围，被函告后仍不停止，构成不正当竞争。

#### 5.1.3 裁判结果

北京第一中级人民法院 2014 年 8 月 7 日宣判，北京奇虎科技有限公司（360）在推出搜索引擎的伊始阶段，没有遵守百度的爬虫协议，违反了《反不正当竞争法》相关规定，应赔偿百度经济损失及合理支出共计 70 万元。

### 5.2 大众点评诉百度不正当竞争案

#### 5.2.1 主要事实

大众点评网收集了大量商户信息，并吸引大量消费者通过真实体验发布点评信息。大众点评网中的用户点评等内容已经成为广大消费者选择相关商家和服务的重要参考资料，取得了良好的社会效益和经济效益。虽然百度爬取大众点评网的点评信息未违反大众点评上的爬虫协议，但百度未经许可，在百度地图、百度知道中大量抄袭、复制大众点评网的点评信息，直接替代了大众点评网向用户提供内容。特别是百度地图，虽然其设置了指向大众点评网的链接,但由于每一条点评信息都是完整的,用户并不需要再去大众点评网查看该信息，导致大众点评网的流量减少。此外，百度地图在大量使用大众点评网点评信息的同时,又推介自己的团购等业务,攫取了大众点评网的部分交易机会。

#### 5.2.2 裁判要点

在靠自身用户无法获取足够点评信息的情况下,百度公司通过技术手段,从大众点评网等网站获取点评信息,用于充实自己的百度地

图和百度知道。百度公司此种使用方式,实质替代大众点评网向用户提供信息,对汉涛公司造成损害。百度公司并未对于大众点评网中的点评信息作出贡献,却在百度地图和百度知道中大量使用了这些点评信息,其行为具有明显的”搭便车”、”不劳而获”的特点。因此,百度公司大量、全文使用涉案点评信息的行为违反了公认的商业道德和诚实信用原则,具有不正当性,给汉涛公司造成了实质损害,构成不正当竞争。

法院认为,网站通过爬虫协议可以告诉搜索引擎哪些内容可以抓取,哪些内容不能抓取。由于爬虫协议是互联网行业普遍遵守的规则,故搜索引擎违反爬虫协议抓取网站的内容,可能会被认定为违背公认的商业道德,从而构成不正当竞争。但并不能因此认为,搜索引擎只要遵守爬虫协议就一定不构成不正当竞争。爬虫协议只涉及搜索引擎抓取网站信息的行为是否符合公认的行业准则的问题,不能解决搜索引擎抓取网站信息后的使用行为是否合法的问题。 本案中,百度公司的搜索引擎抓取涉案信息并不违反爬虫协议,但这并不意味着百度公司可以任意使用上述信息,百度公司应当本着诚实信用的原则和公认的商业道德,合理控制来源于其他网站信息的使用范围和方式。百度公司拥有强大的技术能力及领先的市场地位,若不对百度公司使用其他网站信息的方式依法进行合理规制,其完全可以凭借技术优势和市场地位,以极低的成本攫取其他网站的成果,达到排挤竞争对手的目的。

### 5.2.3 裁判结果

上海知识产权法院二审维持上海市浦东新区人民法院一审原判，百度立即停止以不正当的方式使用大众点评网的点评信息，同时赔偿大众点评网经济损失及为制止不正当竞争行为所支付的合理费用

## 5.3 新浪微博诉脉脉非法抓取用户信息不正当竞争案

### 5.3.1 主要事实

2014 年 8 月，微博方面发现在脉脉产品内，大量非脉脉用户直接显示有新浪微博用户头像、名称、职业和教育等信息。即可能在从未通过微博登录脉脉的情况下，脉脉上仍能够直接搜索到用户的个人信息。而在此之前，微博和脉脉一直有合作，用户可以通过微博账号和个人手机号注册登录脉脉（OpenAPI），用户注册时还要向脉脉上传个人手机通讯录联系人。但在上述时间发生后，双方终止合作，新浪也据此向脉脉提起诉讼，认为被告非法抓取、使用微博用户信息等。

### 5.3.2 裁判要点

脉脉获取新浪微博用户信息的行为是否合法正当。此案涉及互联网环境下用户信息的获取和使用，在案件查明和适用法律层面判断淘友公司是否构成不正当竞争时，要判断其获取、使用新浪微博平台用户信息等行为是否具有合法性和正当性。关于合法性，此案的争议焦点在于脉脉软件的行为是否符合《开发者协议》的约定。而脉脉软件未能对在合作结束后仍使用新浪微博用户的用户信息之必要性给予合理解释，则说明其行为不具合法性。

关于正当性，即其行为是否符合行业管理等正当使用之目的，可

以归纳为：如何判断脉脉注册用户收集通讯录联系人手机号与新浪微博用户信息形成对应关系的正当性。脉脉软件承认这种对应关系的展示是为了引导脉脉用户邀请新浪微博用户加入脉脉，该行为显然属于为脉脉软件增加用户规模的市场行为，非必要的功能性设置，因而其行为不具有正当性。

#### 5.3.3 裁判结果

脉脉软件获取新浪微博信息的行为存在主观过错，违背了第三方通过 Open API 获取用户信息时应坚持“用户授权”+“平台授权”+“用户授权”的三重授权原则，违反了诚实信用原则和互联网中的商业道德，故其行为构成不正当竞争。判令脉脉立即停止涉案不正当竞争行为，并赔偿新浪微博 200 万元。

### 5.4 淘宝诉美景大数据产品不正当竞争纠纷案

#### 5.4.1 主要事实

淘宝公司系“生意参谋”零售电商数据产品的开发者和运营者，该数据产品主要为淘宝、天猫商家的网店运营提供数据化参考服务、帮助商家提高经营水平，淘宝公司对该数据产品享有竞争性财产权益。美景公司运营其“咕咕生意参谋众筹”网站，以提供远程登录服务的方式，招揽、组织、帮助他人获取“生意参谋”数据产品中的数据内容，并从中获取利益。

美景公司开发了名为“咕咕互助平台”与“咕咕生意参谋众筹”的软件与平台，让已订购淘宝生意参谋的用户可以通过咕咕互助平台分享、共用其子账户。通过分享、共用子账户，已订购生意的用户可

以获得佣金。美景公司还提供为远程登陆提供技术支持，而美景公司自然也从生意参谋的有偿共享过程中获得分成。

淘宝公司认为美景公司的行为构成了对生意参谋产品的实质性替代，“直接导致了淘宝公司数据产品订购量和销售额的减少，极大损害了淘宝公司的经济利益，同视恶意破坏了淘宝公司的商业模式，严重扰乱了大数据行业的竞争秩序，已构成不正当竞争行为。”

### 5.4.2 裁判要点

（1）淘宝对数据产品是否具有法定权益

法院认为：涉案“生意参谋”数据产品所提供的数据内容虽然来源于原始用户信息数据，但经过淘宝公司的深度开发已不同于普通的网络数据，其提供的数据内容虽然同样源于网络用户信息，但经过网络运营者大量的智力劳动成果投入，经过深度开发与系统整合，最终呈现给消费者的数据内容，已独立于网络用户信息、原始网络数据之外，是与网络用户信息、原始网络数据无直接对应关系的衍生数据。网络运营者对于其开发的大数据产品，应当享有自己独立的财产性权益。随着互联网科技的迅猛发展，网络大数据产品虽然表现为无形资源，但可以为运营者所实际控制和使用，网络大数据产品应用于市场能为网络运营者带来相应的经济利益。随着网络大数据产品市场价值的日益凸显，网络大数据产品自身已成为了市场交易的对象，已实质性具备了商品的交换价值。对于网络运营者而言，网络大数据产品已成为其拥有的一项重要的财产权益。另一方面，网络数据产品的开发与市场应用已成为当前互联网行业的主要商业模式，是网络运营者市

场竞争优势的重要来源与核心竞争力所在。因而，可确认淘宝对数据产品具有法定权益。

（2）美景公司的行为，是否构成不正当竞争，侵犯淘宝公司的利益

法院认为，美景公司未付出自己的劳动创造，仅是将“生意参谋”数据产品直接作为自己获取商业利益的工具，其所用“生意参谋”数据产品也仅是提供同质化的网络服务。此种拿他人市场成果直接为己所用，从而获取商业利益与竞争优势的行为，明显有悖公认的商业道德，属于不劳而获“搭便车”的不正当竞争行为。

#### 5.4.3 裁判结果

杭州铁路运输法院 2017 年裁定，美景公司停止不正当竞争行为，并赔偿淘宝公司经济损失及合理费用共 200 万元。

### 5.5 央视与视畅公司侵害著作权纠纷、不正当竞争纠纷案

#### 5.5.1 主要事实

央视经国际奥委会授权，在中国境内独家享有通过信息网络，转播中央电视台制作、播出伦敦奥运会开幕式电视节目的权利。视畅公司未经许可，通过其运营的 www.kanketv.com 网站及由其发布的“看客影视”安卓系统客户端软件，向公众实时转播伦敦奥运会开幕式节目，视畅公司采用的方式是利用网络爬虫技术获取了源于 www.cntv.cn 网站中的涉案视频链接并向社会公众提供。

#### 5.5.2 裁判要点

被告在未偿付对价的情况下，利用原告保留权利的商业资源牟利

并致原告利益受损，其行为已构成不正当竞争。首先，凝结创造性技术手段，记录并还原奥运会开幕式全过程所形成的作品—开幕式节目作为一种商业资源更是蕴含着相当程度的市场交易价值。央视国际公司经国际奥委会、中央电视台的独享授权，通过互联网上播放包括开幕式节目在内的伦敦奥运会赛事节目。鉴于原告央视国际公司已多次公开发表声明保留权利，明确表达了其对所占有涉案授权资源权利独享、限制互联网传播的意愿，因而开幕式节目均不应视为可供社会公众任意分享、利用的互联网公共资源。其次，视畅公司以搜索链接的方式传播开幕式节目，是建立在傍附本属原告市场资源的基础之上。视畅公司在无需付出交易成本或付出交易成本甚微的条件下，即可凭借向公众提供与原告实时转播之开幕式节目相同的感官体验，获得与原告视频服务内容一致的竞争优势。同时，本院注意到，视畅公司在涉案网站上多处加载有相应的商业广告。有鉴于此，视畅公司未经授权以搜索链接方式传播开幕式节目的行为，必然会在一定程度上分流本属于原告的用户群体，增加自身的商业广告收益及客户端软件的下载数量。故认定视畅公司的行为已实质性地利用了央视享有权益的市场资源，打破原有的交易秩序，挤占原告的交易机会，并损害其竞争权益。

综上，法院认定，视畅公司所获得的竞争利益，是通过“食人自肥”的不正当手段实现的，其获利核心在于攫夺本属原告的合法商业利益与竞争优势。在损害央视商业利益的同时，视畅公司的行为亦违背了诚实、公平的商业伦理，破坏了原本稳定、有序的竞争秩序，构

成不正当竞争。

### 5.5.3 裁判结果

上海知识产权法院二审维持上海市徐汇区人民法院一审判决，裁定视畅公司赔偿央视经济损失及合理开支。

## 5.6 大众点评网诉爱帮网不正当纠纷案

### 5.6.1 主要事实

大众点评网中的商户介绍和用户点评已经成为广大消费者选择相关商家和服务的重要参考资料，并进而为大众点评网取得了良好的社会效益和经济效益。大众点评网投入大量人力和资金创作了相关商户介绍信息，并享有其全部著作人身权和财产权。根据大众点评网与大众点评网注册用户签订的在线协议，大众点评网独家拥有大众点评网中所载用户点评内容的著作财产权。

爱帮网公司未经原告同意，擅自复制原告享有著作权的商户介绍和点评，刊登于其经营管理的爱帮网上，试图建立“最大、最全的生活信息网上平台”，并在全国范围内进行广告招商活动。

### 5.6.2 裁判要点

爱帮网使用大众点评网用户简介和用户点评，是否构成不正当竞争。大众点评网的商户简介和用户点评，是韩涛公司搜集、整理和运用商业方法吸引用户注册而来。汉涛公司为此付出了人力、财力、物力和时间等经营成本，由此产生的利益应受法律保护。对于大众点评网的用户简介和用户点评，爱帮科技公司未付出劳动、未支出成本、未作出贡献，却直接利用技术手段在爱帮网上展示，并以此获取商业

利益，属于反不正当竞争法理论中典型的“不劳而获”和“搭便车”行为。爱帮科技公司的这一经营模式违反公平原则和诚实信用原则，违反公认的商业道德。与此同时，爱帮网使用的源于大众点评网的内容中，商户简介和大众点评网完全一致，用户点评和大众点评也没有实质性的区别。通过爱帮网，用户可直接获取商户简介的全部内容和用户点评的绝大部分内容，基本实现获取信息的目的。虽然爱帮网注有“在大众点评发表”字样和链接标识，但爱帮网已对全部商户简介内容和绝大部分点评内容进行了充分展示，网络用户一般不会再选择点击大众点评链接标识。因此，爱帮版的商户简介和用户点评已经构成对大众点评网相应内容的实质性替代，必将不合理的损害韩涛公司的商业利益。因此，爱帮网的行为属于不正当竞争。

#### 5.6.3 裁判结果

爱帮科技公司停止在爱帮网使用源自大众点评网的商户简介和用户点评，并赔偿韩涛公司经济损失及诉讼合理开支共计 50 万元。

### 5.7 CraigslistInc. v. 3Taps Inc., 942 F. Supp. 2d 962

#### 5.7.1 主要事实

原告 Craigslist 是一家知名的大型互联网分类广告网站，为用户提供免费的互联网广告发布和浏览服务。被告 3Taps 从原告 Craigslist 处聚合广告信息并进行二次展示。Craigslist 指控 3Taps 直接从 Craigslist 网站上实时复制（或“爬取”）其上面分布的全部内容。3Taps 运营“Craigslist API”来允许第三方网站直接从 Craigslist 上获取信息，同时还运营网站“craiggers.com”，上面

基本上复制了整个 craigslist 网站的内容。

在获悉 3Taps 的爬取行为后，Craigslist 采取了以下两步措施：首先，它向 3Taps 发送了一封制止函（cease and desist letter），告知后者及其代理人、员工、关联公司等不再有权访问 Craigslist 网站，并且禁止他们再以任何理由访问 craigslist 的网站或服务。其次，Craigslist 更改了其网站的设置，对与 3Taps 相关的网址设置了 IP 壁垒（即不允许 3Taps 的相关 IP 网段访问 Craigslist）。但 3Taps 通过使用不同的 IP 地址和代理服务器来隐藏其身份绕过 IP 壁垒并继续爬取 Craigslist 数据。

### 5.7.2 裁判要点

3Taps 在访问 Craigslist 并有意从“受保护的计算机系统”中获取信息的行为是否属于“未经授权”。法院认为：当 Craigslist 撤销 3Taps 对其网站的访问授权后，3Taps 仍然继续访问并爬取数据的行为属于“未经授权”。是否具有“授权”依赖于授予或禁止访问的“授权方”的决定。在本案中，授权方是 Craigslist。 Craigslist 首先授予赋予了公众访问其公开网站的授权。然后，其作为授权方又取消了对 3Taps 的授权——通过明确告知 3Taps 不得“出于任何原因”访问其网站，且设置了技术壁垒以切断相关访问。因此 3Taps 在授权被取消后继续访问的行为便属于“未经授权”。

### 5.7.3 裁判结果

3Taps 希望法院认可“公共网站的所有者无权撤销特定用户访问该网站的授权”的请求，法院不予支持。

## 5.8 EarthCam,Inc. v. OxBlueCorp., 703 Fed.Appx.803, 808

### 5.8.1 主要事实

EC 和 OxBlue 都是经营影像器材和解决方案的公司。EC 的一个用户将自己账户密码给了 OxBlue 公司，希望经营类似业务的 OxBlue 能帮忙解决一些技术问题。后者登录了该账户密码，并对 EC 社群论坛上的大量图片等其他信息进行抓取。

### 5.8.2 裁判要点

法官认为，虽然 CFAA 并没有明确规定用户不得与他人共享账户信息，甚至用户都可以把账户密码在网上公开都是 ok 的，但是 EC 网站上明确声明如果用户将账户信息给他人使用，违反了其“使用条款”。这属于 CFAA 认定的”超出权限“——网站只授权给当事人使用，其他人用当然超出了权限。也就是说，如果违反了有关涉案计算机的任何政策或使用条款（EULA），可以被认定为“超出了授权访问权限”。

### 5.8.3 裁判结果

EC 公司胜诉。

## 5.9 hiQLabs, Inc. v. LinkedIn Corp., 273 F. Supp.3d 1099

### 5.9.1 主要事实

hiQ Lab 公司的主营业务就是利用爬虫从 LinkedIn 网站上获取公开的求职者数据，帮助企业分析和管理人力资源。招致 LinkedIn 的不满并采取技术反制，但这一经营模式招致 LinkedIn 的不满，LinkedIn 拒绝 HiQ 使用其职场数据并采取技术反制。而这些数据正

是 HiQ 公司核心业务开展的基础，禁止数据使用将给 HiQ 公司带来毁灭性的打击。基于此，hiQ 将 LinkedIn 诉至法院。

### 5.9.2 裁判要点

法院认为，第一，在网站上公开的信息不是 CFAA 法条中阐明的“受保护的计算机”，所以没有违反 CFAA，将对 CFAA 的解释重心从传统的“未经授权”延伸至“未经允许访问受保护的计算机”。第二，根据加州反不正当竞争法（California’sUnfair Competition Law），LinkedIn 将它在在职场社交领域的竞争优势转移到职场数据分析领域，是不正当的竞争行为。第三，美国最高法院最新裁定社交媒体类似于一个“现代公共广场”，用户在 LinkedIn 上的信息相当于公共场所言论，由此根据加州宪法对言论自由权的保护，hiQ 声称 LinkedIn 不能限制别的公司去获取这些相当于“言论”的信息，应当支持。

### 5.9.3 裁判结果

hiQ 公司胜诉。