SYMBIOTIC RELATIONSHIPS: PRAGMATIC ACCEPTANCE OF DATA SCRAPING

Jeffrey Kenneth Hirschey†

Automated data collection on the Internet is nothing new, and scrapers continually access and repost data for other websites. Although past scrapers parasitically reposted information to directly *compete* with the scraped website, scrapers now offer mutualistic benefits that can *help* scraped websites. This information is often highly valuable to the businesses that collect it, and they go to great lengths to protect it. Search engines, PageRank,[1] and advertising all use bots to collect information stored by others. With the advent of improved data analytics and the increased technical ease of gathering data, the potential benefits of scraping data have never been higher. This increasingly complex symbiotic relationship between scrapers and data hosts, from parasitic to mutualistic, overlays the already uncertain legal background of scraping case law.

Web services can gather information from data hosts—websites that store or house target data—primarily by parsing or scraping data. Parsing generally refers to the collection of information from the data host directly.[2] Parsing accesses a website's underlying data structures through a series of formalized data requests, often through application programming interfaces ("APIs"). PHP and other common server-side scripting languages are perhaps the most widespread parsers on the Internet.[3] In contrast, data scraping can be broadly defined as a data collection technique where a computer program extracts and reposts data from a user output.[4] Data

† J.D. Candidate, 2015, University of California, Berkeley, School of Law.

1. *PageRank*, GOOGLE, https://support.google.com/toolbar/answer/79837?hl=en (last visited Mar. 4, 2014).

2. *See* Cory Janssen, *Definition of Parsing*, TECHNOPEDIA, http://www.techopedia.com/definition/3853/parse (last visited Jan. 30, 2014).

3. PHP is the largest server-side programming language and is used on over 244 million websites and 2.1 million web servers. *See History of PHP and Related Projects*, PHP, http://us1.php.net/history (last visited Jan. 30, 2014); *see also Usage Stats for January 2013*, PHP, http://www.php.net/usage.php (last visited Jan. 30, 2014).

4. *See Scraping*, PC MAG., http://www.pcmag.com/encyclopedia/term/57344/scraping (last visited Jan. 30, 2014); John Wagnon, *Web Scraping—Data Collection or Illegal Activity*, DEVCENTRAL (May 16, 2013), https://devcentral.f5.com/articles/web-scraping-data-collection-or-illegal-activity#.UovBOGSglX8.

scraping typically collects data from screen outputs or extracts data from the HyperText Markup Language ("HTML") code that most websites display.[5] Although parsing provides stable access to underlying web data, scraping can access large amounts of data without the permission, or even knowledge, of the data host. Although the technical distinctions between scraping and parsing remain murky, scraping is often regarded with a more negative connotation.[6]

Many widespread uses of data scraping are mutualistic: they benefit both data hosts and scrapers. Scraping services allow many users to find the information they seek more easily. Familiar web search engines are essentially scrapers that pull small amounts of data—i.e., the search terms a user enters—to link a user to relevant webpage results.[7] Yet search engines have avoided much of the negative stigma associated with scraping, and they are an instrumental part of the online ecosystem. Indeed, most search engines are instead referred to as "web indexers" or "web crawlers" even though the URL and hyperlink data collected can often be directly used to perform data scraping.[8] Data scraping may also be used to track the way webpages link to each other. Google's ubiquitous PageRank algorithm is perhaps the largest scraping system and uses a web crawler called GoogleBot to scrape data from billions of webpages.[9] This model is predicated upon unfettered access to data, and data hosts provide little resistance given the overwhelming benefit that they receive.

Despite some positive applications of scraping, scraping is parasitic when scrapers benefit at the exclusion or detriment of data hosts. Scrapers can collect information without the consent of data hosts and may undercut a website's revenue by republishing scraped data without requiring users to

5. Friedrich Lindenberg, *Getting Data from the Web*, DATA JOURNALISM HANDBOOK, http://datajournalismhandbook.org/1.0/en/getting_data_3.html (last visited Jan. 30, 2014).

6. Arpan, *Data Scraping vs. Data Crawling*, PROMPT CLOUD (May 30, 2012), http://blog.promptcloud.com/2012/05/data-scraping-vs-data-crawling.html; *see also Crawler vs scraper*, STACK OVERFLOW, http://stackoverflow.com/questions/3207418/crawler-vs-scraper (last visited Jan. 30, 2014).

7. For general background on information retrieval on the web, see Mei Kobayashi & Koichi Takeda, *Information Retrieval on the Web*, 32 ACM COMPUTING SURVEYS 144 (2000), *available at* http://dl.acm.org/citation.cfm?doid=358923.358934.

8. *What Is a Screen Scraper?*, WISEGEEK, http://www.wisegeek.com/what-is-a-screen-scraper.htm (last visited Jan. 30, 2014).

9. GoogleBot is the script that gathers data to support Google's search engine. Google describes the action of the bot stating, "We use a huge set of computers to fetch (or 'crawl') billions of pages on the web." *See GoogleBot*, GOOGLE, https://support.google.com/webmasters/answer/182072?hl=en (last visited Jan. 30, 2014). *See generally How Google Search Works*, GOOGLE, http://www.google.com/competition/howgooglesearchworks.html (last visited Jan. 30, 2014).

view supporting advertisements. Worse still, scrapers may derive their own ad revenues, viewers, and customers by taking content directly from another data host.[10] Scraping may collect personally identifying information ("PII") thought to be private and can have serious privacy implications.[11] Scraping activity may even directly harm a data host's core services or prevent users from using those services.[12]

Unsurprisingly, data hosts have fought to control their data using various legal and technological methods in myriad contexts online. In perhaps the most famous data scraping case, online auction website eBay sued the now defunct auction compiler Bidder's Edge over republication of auction data from eBay.[13] Airline price aggregators, websites that aggregate and display price and flight information from multiple airline carriers, such as Kayak, Orbitz, and Expedia, have been subject to legal action.[14] Real estate multiple listing services ("MLS") have sued data scrapers that reposted real estate advertisements, descriptions, and listing photos without obtaining consent.[15]

---

10. Google AdSense has tried to limit monetization of scraped content through AdSense's terms of service and program policies. The policies strictly limit the permitted uses of scraped data. *See AdSense Program Policies*, GOOGLE, https://support.google.com/adsense/answer/48182?hl=en&ref_topic=2864301 (last updated Jan. 10, 2014).

11. Scrapers have targeted sensitive information message boards where people discussed emotional disorders, collected PII for background checks, shared private information from social media networks, and provided data so employers can screen job candidates. *See* Julia Angwin & Steve Stecklow, *'Scrapers' Dig Deep for Data on Web*, WALL ST. J. ONLINE (Oct. 12, 2010, 12:01 AM), http://online.wsj.com/news/articles/SB10001424052748703358504575544381288117888.

12. Some courts allow data hosts to redress these harms through contract claims, for example when a bot has violated the scraped site's terms of use, as well as a variety of tort claims. In *MDY Industries v. Blizzard*, the Ninth Circuit found that a bot used to automate World of Warcraft gameplay violated an "effective access control measure," triggering a number of causes of actions. MDY Indus., LLC v. Blizzard Entm't, Inc., 629 F.3d 928, 954 (9th Cir. 2010). Use of the bot not only violated Blizzard's Terms of Service, but it also gave rise to a claim for tortious interference and violated the Digital Millennial Copyright Act ("DMCA"), 17 U.S.C. § 1201(a)(2). *Id.* Further, a company that sold "automated devices to access and navigate through Ticketmaster's website," allowing users to purchase desirable tickets before other customers, was found by a California district court to violate Ticketmaster's terms of use and support several other state and federal claims. *See* Ticketmaster L.L.C. v. RMG Techs., Inc., 507 F. Supp. 2d 1096, 1102 (C.D. Cal. 2007).

13. eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058 (N.D. Cal. 2000).

14. Southwest sued Orbitz alleging Orbitz had posted false and misleading price information about Southwest's fare information. *See* Sw. Airlines v. Orbitz LLC, No. 2:01-cv-04068 (C.D. Cal. filed May 3, 2001); *see also* Michael Mahoney, *Orbitz Sued by Southwest Airlines*, E-COMMERCE TIMES (May 4, 2001, 10:01 PM), http://www.ecommercetimes.com/story/9518.html.

15. *See* Metro. Reg'l Info. Sys., Inc. v. Am. Home Realty Network, Inc., 722 F.3d 591, 592 (4th Cir. 2013). *See generally* Marianne M. Jennings, *Multiple Listing Services—Antitrust and Policy*, 32 REAL EST. L.J. 140 (2003).

Many major banks and financial institutions have sued financial data aggregators, financial money management applications such as Mint, or account aggregation vendors.[16] Social media giant Facebook has sued third-party applications that have attempted to access and republish Facebook's user data.[17] Financial brokerage houses have sought to stem the flood of investing advice leaking from their websites by suing those who republish and threatening to break the established price-discrimination model.[18] Data hosts have even sued scrapers with whom they initially contracted to catalog their data.[19] There are countless examples of recent cases where data hosts sought legal remedies for the collection and dissemination of their data.

Recently, Craigslist has been one of the most vigilant, and visible, data hosts to take a stand against scraping. With over fifty billion page views per month, Craigslist is the third most visited American Internet company.[20] Craigslist operates a series of online classified advertisements where users post hundreds of millions of ads for goods and services each year.[21] Recently, Craigslist sued 3Taps, PadMapper, and Lovely, which are services that sought to augment Craigslist's interface by providing users with an integrated, easy-to-navigate map that displayed the locations of user-generated ads.[22] At the

16. For a brief overview of financial and account aggregators, see generally Nathan J. Sult, *"Show Me the Money": The Emerging Technology of Internet Financial Aggregation*, 5 HAW. B.J. 20 (Mar. 2001); *see also* Kimberly L. Wierzel, *If You Can't Beat Them, Join Them: Data Aggregators and Financial Institutions*, 5 N.C. BANKING INST. 457 (2001).

17. *See, e.g.*, Facebook, Inc. v. Power Ventures, Inc., 844 F. Supp. 2d 1025, 1027 (N.D. Cal. 2012).

18. Price discrimination models charge different rates for similar goods in different markets. These models require a separation between markets or the goods will simply drop to the lowest price across any of the markets. In an online information context, price discrimination often involves selling information that is not widely available or publically accessible. If this information becomes public then there is no longer an incentive to pay for it, and the data host loses customers. Brokerage firms have attempted to protect the dissemination of early financial recommendations using a variety of misappropriation and copyright claims. *See, e.g.*, Barclays Capital Inc. v. Theflyonthewall.com, Inc., 650 F.3d 876, 878 (2d Cir. 2011).

19. Scraping violations can be alleged contractually when a data host contracts directly for a data scraping service. Breach can occur if this contract is then violated and, unlike other scraping contexts, the scraper can also sue the data host. Data hosts have even been sued when they hire vendors to scrape their own data to create product catalogues. *See* Edgenet, Inc. v. Home Depot U.S.A., Inc., 658 F.3d 662, 663 (7th Cir. 2011); Snap-on Bus. Solutions Inc. v. O'Neil & Assocs., Inc., 708 F. Supp. 2d 669, 671–72 (N.D. Ohio 2010).

20. *Craigslist Factsheet*, CRAIGSLIST, http://www.craigslist.org/about/factsheet (last visited Mar. 7, 2013); Craigslist, Inc. v. 3Taps, Inc., 942 F. Supp. 2d 962 (N.D. Cal Apr. 30, 2013).

21. *Craigslist*, 942 F. Supp. 2d at 966.

22. *See, e.g.*, *Terms of Service—§ 1.2: 3taps' Manifesto*, 3TAPS, http://www.3taps.com/terms.php (last updated Oct. 22, 2010).

time of writing, this dispute is currently ongoing in the Northern District of California, with all of Craigslist's major claims having survived a motion to dismiss.[23] Although still incipient, this litigation represents one of the most recent high-profile scraping cases and provides a good model to analyze the legal claims commonly asserted by data hosts against scrapers.[24] Understanding the legal standards for these claims is critical to understanding the legal framework that surrounds suits between data hosts and scrapers.

Despite Craigslist's hard-line stance against data scrapers, this Note argues that data hosts stand to benefit when their users can most effectively access the data they seek—even when scrapers, rather than the data hosts themselves, facilitate this access. Data hosts should recognize the benefit scrapers can provide and take a pragmatic approach to those who scrape their data. Specifically, data hosts should only seek legal remedies against scrapers when (1) the scraper presents a threat to the data host's core business and (2) the data host has a strong enough claim to prevail legally against the scraper.

This Note uses *Craigslist v. 3taps* to discuss the current legal regime surrounding data scraping on the Internet. This Note intends to be of practical significance to the data host community, both in terms of data management and litigation strategy, and in helping them recognize new opportunities to use scraping to their advantage. Part I sketches the technical background on scraping, discusses the current regime of legal protections for databases, outlines the key legal claims often brought against scrapers, and lays out the common legal and technological steps that data hosts have taken to protect their data. Part II examines recent scraping litigation to frame the contexts when suits have been successful in protecting data hosts and user interests, and when suits have been largely unsuccessful to prevent new services from supplanting existing data hosts. Part II also focuses on what factors a data host should consider before bringing suit against scrapers and offers suggestions for how data hosts might establish frameworks to work more cooperatively with scrapers. The Note concludes with a brief consideration of how changing cyberlaw regimes may affect a data host's ability to bring suit against scrapers in the near future.

23. *Craigslist*, 942 F. Supp. 2d at 965–66.

24. Three of Craigslist's claims—a Computer Fraud and Abuse Act ("CFAA") claim, a copyright infringement claim, and a trespass claim—survived the motion to dismiss. *Id.* These are typical claims exerted by data hosts. Although early in the litigation, surviving the motion to dismiss suggests these claims are not being brought frivolously and that they may have some bite.

I. SCRAPING BACKGROUND AND HOW DATA HOSTS CAN PROTECT THEIR DATA

To understand the legal and business implications of data scraping, it is important to first understand how scraping works from a technical perspective. Section I.A briefly discusses the Craigslist litigation, while Section I.B overviews the technical foundations of scraping and how websites can be scraped online. Section I.C next discusses the current legal protections available for databases and their implication in scraping cases. Section I.C also examines the current legal standard for the major claims typically asserted by data hosts against scrapers: violation of the Computer Fraud and Abuse Act ("CFAA"), trespass to chattel, and compilation copyright claims. Lastly, Section I.D outlines the most common defensive measures, both legal and technical, that data hosts take against scrapers.

A. OVERVIEW OF THE CRAIGSLIST LITIGATION

Craigslist's claims against 3Taps, Padmapper, and Lovely represent only the most recent case in a series of ongoing lawsuits against scrapers by Craigslist. Craigslist has built enormous market share in online classified advertisements and actively seeks to prevent competitors from entering the market by pursuing legal action against nearly every service that has scraped it.[25] In 2005, Craigslist stopped Oodle, a service that used screen-scraping to display ads in a search engine.[26] In 2007, Craigslist blocked Listpic, a service that displayed photos with user ads.[27] In 2009, Craigslist blocked Flippity, a service that displayed Craigslist postings within a map.[28] Each of these services sought to improve Craigslist's underlying functionality by providing users with additional services or features beyond the native Craigslist interface. Despite these potential benefits to user experiences, Craigslist has

25. *See* John Koetsier, *3Taps Sues Craigslist to Save the Internet (No Seriously)*, VENTURE BEAT (Sept. 24, 2012, 4:52 PM), http://venturebeat.com/2012/09/24/3taps-suing-craigslist-save-internet.

26. Craigslist continued to block Oodle even when the service discontinued its use of screen scraping to collect data from Craigslist. *See* John Battelle, *Craigslist Blocks Oodle*, JOHN BATTELLE'S SEARCH BLOG (Oct. 14, 2005), http://battellemedia.com/archives/2005/10/craigslist_blocks_oodle.php.

27. *See* Meg Marco, *Craigslist Blocks "ListPic" Tool for Viewing Craigslist Pictures*, CONSUMERIST (June 8, 2007), http://consumerist.com/2007/06/08/craigslist-blocks-listpic-tool-for-viewing-craiglist-pictures; John Musser, *Craigslist Blocks Mashup Listpic*, PROGRAMMABLE WEB (June 18, 2007), http://blog.programmableweb.com/2007/06/18/craigslist-blocks-mashup-listpic.

28. *See* Jason Kincaid, *Craigslist Blocks Yahoo Pipes After Dev Shows Craig His New Mashup*, TECHCRUNCH (Dec. 1, 2009), http:/techcrunch.com/2009/12/01/craigslist-yahoo-pipes-flippity.

consistently resorted to threats of or actual litigation to prevent unauthorized use of its data.

In the current litigation, Craigslist alleges that Padmapper, 3Taps, and Lovely improperly gathered classified ad information from Craigslist and reposted that information on their own websites alongside a map interface that plotted the location of the user-generated classified ads.[29] Interestingly, Craigslist itself now offers a similar mapping service.[30] The district court for the Northern District of California recently denied the defendants' motion to dismiss the compilation copyright infringement, violation of the Computer Fraud and Abuse Act ("CFAA"), and trespass to chattel claims.[31] These are three of the most commonly asserted claims brought by data hosts against scrapers, and all three survived the motion to dismiss. Despite the early state of the Craigslist litigation, these claims remain the key legal claims in most scraping cases.

The Craigslist litigation is an interesting comparison tool for other scraping cases. Despite changing legal regimes,[32] the three major claims in past scraping cases are the claims that Craigslist exerted here. These claims mostly rely upon the protections that the data host set up rather than inherent protections for data generally.

B. TECHNOLOGICAL BACKGROUND FOR DATA SCRAPING

When users access the Internet, they are greeted with visual representations of underlying web data. When a user types in a website's domain name, for example, www.google.com, a protocol called a Domain

29. Craigslist outlined the technological specifics of the 3Taps, Padmapper, and Lovely services in its complaint. *See* Complaint at 8–15, Craigslist, Inc. v. 3Taps, Inc., 942 F. Supp. 2d 962 (N.D. Cal 2013) (No. CV-12-3816-LB).

30. After years of user demand for a mapping function, Craigslist finally added a mapping function in 2012. This mapping function is very similar to the options provided by scrapers in the past. *See* Josh Ong, *Craigslist Rolls Out New Map View Feature for Apartment Searches*, THE NEXT WEB (Oct. 4, 2012), http://thenextweb.com/insider/2012/10/04/craigslist-rolls-out-new-map-view-feature-for-apartment-searches.

31. The district court did note that it was still very early in the litigation and many issues presented by these claims, such as a demonstration of harm required under trespass to chattel claims, would still need to be demonstrated at a later stage in the litigation. *See Craigslist*, 942 F. Supp. 2d. at 966. The district court also severed the defendants' antitrust counterclaims, and it is likely the court will only address them following the resolution of Craigslist's claims. *See id.* at 982. Although antitrust discussions are beyond the scope of this Note, it is interesting that antitrust counterclaims are commonly asserted against data hosts by scrapers. It is presently unclear what level of data control would have significant antitrust implications thought it is easy to hypothesize a data host denying accesses to data would likely trigger antitrust concerns at some point.

32. *See infra* Section I.C.

Name Service ("DNS") translates that domain name into an Internet Protocol ("IP") address, which the user's web browser can then access to display that website's or server's data.[33] Users can also navigate to websites by following Uniform Resource Locators ("URLs") that link different websites together.[34] To present data in a user-friendly manner, web browsers create graphic representations of websites from HTML code and data that the website supplies.[35]

Scraping accesses data either via data displayed to a user's screen or from the underlying HTML code. Because these are designed as user outputs, it is technically simple to pull data from them. A quick web search offers numerous options to scrape data: how-to guides about scraping, guidance in writing your own scraping program, and even options to purchase scraping software.[36] Given the ease of access and low barrier to entry for data scraping, it can be difficult to anticipate and prevent the many ways data scraping can be performed. Generally, websites focus on ease of use and accessibility to users, making it easier for scrapers to harvest data from nicely formatted user outputs.

---

33. Although IP addresses were initially introduced in 1987, the current IP—IPv6—was created by the Internet Engineering Task Force and uses 128 bits to store addresses. *See* S. DEERING & R. HINDEN, INTERNET ENG'G TASK FORCE, INTERNET PROTOCOL, VERSION 6 (IPV6): SPECIFICATION (Dec. 1998), http://www.ietf.org/rfc/rfc2460.txt. IPv6 was adopted to increase the total number of available addresses and allow for more efficient routing. *Id. See generally* P. MOCKAPETRIS, INTERNET ENG'G TASK FORCE, DOMAIN NAMES—CONCEPTS AND FACILITIES (Nov. 1987), http://tools.ietf.org/html/rfc1034. The Internet Assigned Numbers Authority ("IANA") administers and allocates IP address and DNS globally. *See About IANA*, INTERNET ASSIGNED NUMBERS AUTHORITY, http://www.iana.org/about (last visited Jan. 30, 2014).

34. Links and URLs are managed by standards set by the World Wide Web Consortium ("W3C"). *See URL: Living Standard*, WHATWG, http://url.spec.whatwg.org (last updated Feb. 3, 2014); *see also* Tim Berners-Lee, *Uniform Resource Locator*, W3C, http://www.w3.org/Addressing/URL/url-spec.html (last updated Nov. 1993). *See generally* Dan Connolly, *Naming and Addressing: URIs, URL, . . .* , W3C, http://www.w3.org/Addressing (last updated Feb. 27, 2006).

35. The W3C maintains a list of HTML tags that browsers use to display webpages. *See HTML Tags*, W3C, http://www.w3.org/History/19921103-hypertext/hypertext/WWW/MarkUp/Tags.html (last visited Jan. 30, 2014).

36. Web search results for scraping yield informative blog posts, python and PHP programming help for scrapers, software available for purchase, and how-to guides aimed at the technically unsavy. *See, e.g.*, SCRAPY, http://scrapy.org (last visited Jan. 30, 2014). *See generally* HARTLEY BRODY, THE ULTIMATE GUIDE TO WEB SCRAPING (2013), *available at* http://blog.hartleybrody.com/web-scraping-guide; Michelle Minkoff, *How to Scrape Websites for Data Without Programming Skills*, POYNTER (May 11, 2010, 7:09 PM), http://www.poynter.org/how-tos/digital-strategies/e-media-tidbits/102589/how-to-scrape-websites-for-data-without-programming-skills.

Although scraping accesses user outputs for data, web data can also be collected in other ways. Web servers interact with each other using PHP code, and this code can be parsed to write data pull requests directly to servers.[37] Web crawling relies upon parsing to access and sort many webpages. Because parsing can be conceptualized as a more formal data request, there are some protocols, robots.txt being perhaps the most popular and widely used, that govern parsing requests.[38] These protocols are pieces of code embedded in webpages that tell crawlers which portions of the web page should or should not be accessed.[39] Yet these protocols are voluntary, and unscrupulous web crawlers can simply ignore them. Further, if the protocols are not properly set up, it is possible that even a well-meaning web crawler may not detect them.[40] Data can also be collected from web indexers that follow URLs and determine how webpages are connected to each other.[41] Lastly, data can be collected from APIs. APIs are code interfaces that allow programmers to make very formal data requests from websites within a specific interface.[42] These interfaces vary from service to service, but

37. Even basic programming tutorials teach how to use PHP to process data pull requests from servers. *See* Daniel Shiffman, *Tutorials: External Data into Processing II*, LEARNING PROCESSING, http://www.learningprocessing.com/tutorials/external-data-into-processing-2 (last visited Jan. 30, 2014).

38. The Robots.txt protocol was established as a voluntary protocol to allow websites to dictate their preferences in what portions of a site bots should and should not access. *See* WEB ROBOTS PAGES, http://www.robotstxt.org (last visited Jan. 30, 2014).

39. Robots.txt protocols are small pieces of code placed in the header of websites to signal to bots. When a bot retrieves data from a website it encounters, the bot can then follow the directions provided by the robots.txt protocol. *See About /robots.txt*, WEB ROBOTS PAGES, http://www.robotstxt.org/robotstxt.html (last visited Jan. 30, 2014).

40. Websites that do not implement protocols correctly have been less successful in arguing that they intended to protect their data. In one case, a scraper copied dog breed data from a data host and reposted it on his own website. *See* Tamburo v. Dworkin, 601 F.3d 693 (7th Cir. 2010). The scraped website contained an improperly set up robots.txt protocol that the web crawler ignored. *Id.* The court found that the website's failure to properly execute the robots.txt protocol did not amount to consent for the website to be scraped, but this protocol had not protected the website's data at all. *Id.*; *see also* Venkat Balasubramani, *Calling Out Scraper for "Stealing" Data Is Not Defamatory*, TECH. & MARKETING L. BLOG (Oct 4, 2013), http://blog.ericgoldman.org/archives/2013/10/calling_out_scr.htm.

41. Many web indexers work similarly: they rely upon an efficient indexing of websites through use of a bot, and then use that index to provide faster search results or perform data analysis. *See generally* Sergey Brin & Lawrence Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, 30 COMPUTER NETWORKS & ISDN SYS. 107 (1998), *available at* http://infolab.stanford.edu/~backrub/google.html.

42. Although APIs may perform differently, they are all designed to interface between programs. Perhaps the most widely used APIs today are those of Google and YouTube. *See API Overview Guide*, GOOGLE, https://developers.google.com/youtube/getting_started (last updated May 10, 2013).

generally websites create specific methods that programmers can use to access data.[43] APIs are perhaps the most robust way web data can be accessed, but the highly structured interfaces and commands set by the data host may greatly limit the type, content, and volume of data that can be accessed.

Scraping ignores these protocols and accesses data directly from user outputs. By ignoring the underlying formats and data structures that store web data, scrapers can collect vast amounts of data without the permission of the data host.[44] Yet there are downsides to accessing data via scraping. Because scrapers collect data directly from output displays, the success and accuracy of scraping often depends on a website's output display remaining static.[45] Even small changes in a website's display may disrupt scraping. This instability represents a key tradeoff when scraping to collect data: scrapers can access large amounts of data, but any changes by data hosts can be very disruptive.

C. LEGAL PROTECTIONS FOR DATABASES

In the United States there is no direct legal protection for databases.[46] Although European laws and international treaties have extended legal protection to databases, U.S. law does not directly extend copyright protection to databases.[47] The Berne convention protects collections if "by reason of the selection and arrangement of their contents" they "constitute intellectual creations."[48]

---

43. *See id.*

44. Scrapers threaten loss of consumers, system overload, loss of ad revenue, loss of content, and devaluation. *See* Marino Zini, *Security Zone: Can You Prevent Scraping or Data Harvesting?*, COMPUTERWEEKLY.COM (Nov. 2009), http://www.computerweekly.com/opinion/Security-Zone-can-you-prevent-scraping-or-data-harvesting.

45. The key goal of screen-scraping is to take information that has been formatted to be human-readable and turn it into information that can be processed by a computer. Screen-scraping attempts to find relevant portions of a webpage and manipulate those pieces to extract the data the scraper seeks. *See* Eric Phetteplate, *Web Scraping: Creating APIs Where There Were None*, ACRL TECHCONNECT BLOG (Sept. 30, 2013), http://acrl.ala.org/techconnect/?p=3850.

46. *See* Daniel J. Gervais, *The Protection of Databases*, 82 CHI.-KENT L. REV. 1109 (2007) (examining the flexibility of international norms and their impact upon database protection, as well as the implications of extending protection to purely factual databases).

47. Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299, 33 I.L.M. 1197 (1994).

48. Berne Convention for the Protection of Literary and Artistic Works, Sept. 9, 1886, as revised at Paris on July 24, 1971 and amended in 1979, S. Treaty Doc. No. 99-27 (1986) [The 1979 amended version does not appear in U.N.T.S. or I.L.M.].

In contrast, U.S. patent and copyright law only extends protections to "Science and useful Arts," or creative or inventive works.[49] A database itself only benefits from copyright protection if its organization is sufficiently creative.[50] Practically, copyright protection is minimal when applied to databases.[51] Databases do not typically meet these requirements, and there is no protection for works based solely on the amount of time or effort invested to create them.[52] Websites that seek to compile information thus face free-rider problems; they risk that others can take any factual data they have worked to acquire.[53]

Misappropriation offers some protection for databases, but state law protections for database misappropriation under tort claims are also limited. Nevertheless, data scraping cases have proceeded under state misappropriation theories.[54] Congress has not yet passed federal tort laws protecting the misappropriation of "substantial proportions" of database protections.[55] Several bills have been introduced that would have protected databases built with a "substantial expenditure of financial resources or time," but they have failed to pass in Congress.[56] Although federal tort

49. *See* U.S. CONST. art. I, § 8, cl. 8; 17 U.S.C. § 102 (2012). *See generally* U.S. COPYRIGHT OFFICE, COPYRIGHT LAW OF THE UNITED STATES AND RELATED LAWS CONTAINED IN TITLE 17 OF THE UNITED STATES CODE (2011), *available at* http://www.copyright.gov/title17/circ92.pdf.

50. *See* Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., Inc., 499 U.S. 340 (1991) (holding no copyright protection for facts); Lotus Dev. Corp. v. Borland Int'l, Inc., 49 F.3d 807, 809 (1st Cir. 1995), *aff'd*, 516 U.S. 233 (1996) (holding that there was no copyright protection for computer menu command hierarchy terms).

51. *See* H.R. REP. NO. 106-349, pt. 1, at 10 (1999); *see generally* Julie Wald, Note, *Legislating the Golden Rule: Achieving Comparable Protection Under the European Union Database Directive*, 25 FORDHAM INT'L L.J. 987, 1028 n.186 (2002) (examining the differences between U.S. and E.U. database protection and comparing the E.U. Database Directive with potential U.S. legislation).

52. There is no protection for so-called "sweat of the brow" works, regardless of the amount of energy and time spent to create them. *See Feist*, 499 U.S. at 352–54.

53. *See id.*

54. Tort protections for misappropriation of databases are dependent upon jurisdiction. *See* CollegeSource, Inc. v. AcademyOne, Inc., 653 F.3d 1066 (9th Cir. 2011) (alleging a database misappropriation claim after one website accused another of wholesale database copying).

55. *See* H.R. REP. NO. 106-349, pt. 1, at 11 (1999) (explaining how H.R. 354 would protect database misappropriation). The Collections of Information Antipiracy Act, H.R. 354, 106th Cong. (1999), which failed to pass Congress, would have also protected database misappropriation under a slightly different scheme. *See* Wald, *supra* note 51, at 991, 993–94.

56. Bills introduced in the 108th Congress would have created a federal tort of misappropriation if a "substantial part" of a database had been taken. *See* Database and Collections of Information Misappropriation Act, H.R. 3261, 108th Cong. (2003). Similar

protections for misappropriation might alleviate the free rider problem, there is no current legislation that does so.

In addition to limited protections for the databases themselves, data hosts often house data that intellectual property rights cannot protect. Data hosts may store data that is factual and cannot be protected.[57] Data hosts also often house data that is created by users and not the data host. Much of this user-generated content ("UGC") is simply stored, not owned, by the data host. The content owners, typically users, have the ability to modify their data and may retain the intellectual property and ownership rights of their data unless user agreements or other contractual arrangements modify those rights.[58] Data hosts have less protection for UGC and have faced serious public outcry when attempting to obtain ownership of the underlying rights from the content owners.[59]

To combat the lack of formal doctrinal protections, data hosts have created their own protections through technological and legal restrictions on the access of their information. Much of the data hosts' protection comes from both new causes of action, such as the Computer Fraud and Abuse Act, and adaptations of old causes of action, such as porting trespass claims to the Internet.

1. Computer Fraud and Abuse Act

Data hosts often assert that scrapers have violated provisions of the Computer Fraud and Abuse Act ("CFAA") and its state law counterparts.[60] The CFAA is an antihacking statute that was designed to prevent unauthorized access to websites and servers.[61] There are many state law

attempts have been made in previous bills as well. *See* Collections of Information Antipiracy Act, H.R. 354, 106th Cong. (1999).

57. *See Feist*, 499 U.S. 340.

58. Many websites have terms of use that require users to give exclusive licenses or complete ownership to the website as a condition of signing up. Users may not be aware they are signing over their rights, as it is unlikely many read the terms in great detail, if at all. *See* Curtis Smolar, *Who Owns User-Generated Content?*, VENTURE BEAT (July 12, 2010, 6:00 AM), http://venturebeat.com/2010/07/12/who-owns-user-generated-content.

59. Data hosts have faced serious public-relations backlashes when they have sought exclusive licenses or control over UGC. *See* Mike Masnick, *Craigslist's Abuse of Copyright and the CFAA to Attack Websites that Make Craigslist Better Is a Disgrace*, TECHDIRT (May 1, 2013, 9:29 AM), http://www.techdirt.com/articles/20130501/04342822905/craigslists-abuse-copyright-cfaa-to-attack-websites-that-make-craigslist-better-is-disgrace.shtml.

60. Computer Fraud and Abuse Act ("CFAA"), 18 U.S.C. § 1030 (2012).

61. *See generally* Charles Doyle, CONG. RESEARCH SERV., 97-1025, CYBERCRIME: AN OVERVIEW OF THE FEDERAL COMPUTER FRAUD AND ABUSE STATUTE AND RELATED FEDERAL CRIMINAL LAWS (2010), *available at* http://www.fas.org/sgp/crs/misc/97-1025.pdf.

corollaries to the CFAA, including California Penal Code section 502, which have "functionally identical" requirements to the CFAA.[62] The CFAA imposes criminal penalties on a party who "intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains . . . information from any protected computer."[63]

CFAA violations typically arise only if the scraper has violated an "access restriction" when gathering the data. The language of the CFAA is broad, and courts have determined that nearly any violation of an access restriction to a website may suffice.[64] This includes violations of certain Terms of Use ("TOU"), as well as the evasion of technological defenses, such as IP blocking, that had been used to stop the scraper. Although CFAA claims are normally only exerted against parties that directly scrape a website for data, Craigslist asserted CFAA claims against parties that had gathered Craigslist's data indirectly from a third party.[65]

The Ninth Circuit's opinion in *United States v. Nosal* is a recent interpretation of the language of the CFAA.[66] In *Nosal*, the court found that the phrase "'exceeds authorized access' in the CFAA is limited to violations of restrictions on access to information, and not restrictions on its use."[67] These access restrictions might dictate which users could access data and which data they could access, or might include technological barriers to access.[68] Although the court found that the form of the access restrictions, contractual or technical, was immaterial, the court did find that violations of use restrictions do not violate the CFAA.[69] Although *Nosal* dealt with an employee who had exceeded his authorized access, many data scrapers are outside parties and not employees.[70]

---

62. *See* CAL. PENAL CODE § 502 (West 2011); Craigslist, Inc. v. 3Taps, Inc., 942 F. Supp. 2d 962, 968 (N.D. Cal. 2013).

63. CFAA, 18 U.S.C. § 1030(a)(2), (e)(2) (2012).

64. A definitive interpretation of "access restrictions" in the CFAA continues to be elusive. *See* Facebook, Inc. v. Power Ventures, Inc., 844 F. Supp. 2d 1025, 1035–37 (N.D. Cal. 2012).

65. Craigslist alleged CFAA violations against Lovely and Padmapper, which did not access Craigslist's servers directly. *See Craigslist*, 942 F. Supp. 2d at 971.

66. United States v. Nosal, 676 F.3d 854 (9th Cir. 2012).

67. *Id.* at 863–64.

68. *Id.*

69. *Craigslist*, 942 F. Supp. 2d at 970; *see also Nosal*, 676 F.3d at 862–64.

70. Increasingly, employers are bringing CFAA claims against employees. These claims typically allege that the employee, although an authorized user of the site, has improperly exceeded their access. Generally employees have passwords and do not violate technological restrictions that would stop outside users from entering a corporation's computer systems. *See, e.g.*, People v. Childs, 164 Cal. Rptr. 3d 287 (Ct. App. 2013). *See generally* David J. Rosen,

Even after the court's recent ruling in *Nosal*, it remains unclear exactly what constitutes an access restriction. Many websites employ click-through agreements and TOUs to govern how a user can interact with a website. When a data host suspects or discovers that a violation may have occurred, it traditionally sends cease-and-desist letters to all potential violators. Some have called these cease-and-desist letters "wish lists," and it is uncertain if courts view them as legitimate access restrictions, which would implicate CFAA violations, or merely use restrictions.[71] In the Craigslist litigation, the district court acknowledged the holding in *Nosal* and characterized the cease-and-desist letters as demonstrating "clear statements regarding authorization."[72] The district court did not distinguish the cease-and-desist letters from technological access restrictions that had constituted access restrictions in past cases.[73] Should cease-and-desist letters acquire legal significance (beyond notice) and constitute an access restriction, data hosts will undoubtedly allege more CFAA violations.

CFAA claims remain popular because liability under the CFAA can carry criminal charges,[74] which may serve as a deterrent to scrapers. In the wake of the *Nosal* ruling, however, a court's characterizations of restrictions as access or use restrictions will be a key issue. Under existing case law, data hosts may prevail under the CFAA if they can characterize TOUs, cease-and-desist letters, and click-through agreements as access restrictions. Moving forward, data hosts looking to pursue CFAA claims should make sure that they can adequately control user or scraper access to the website.

2. Compilation Copyright

Data hosts often bring copyright claims against scrapers if underlying content meets copyright requirements. Although these claims do not carry the criminal penalties associated with CFAA violations, copyright claims are popular because the damages can be very large and the copyright protections

Note, *Limiting Employee Liability Under the CFAA: A Code-Based Approach to "Exceeds Authorized Access*," 27 BERKELEY TECH. L.J. 737, 738 (2012).

71. Courts view cease-and-desist letters as legally significant in providing notice to a scraper—the data host thinks they are causing a violation. But these "unregulated wish lists" become much more powerful and potentially disruptive if considered within a court's legal analysis as constituting something beyond notice. *See* Eric Goldman, *Craigslist Wins Routine but Troubling Online Trespass to Chattels Ruling in 3Taps Case (Catch-up Post)*, TECH. & MARKETING L. BLOG (Sept. 20, 2013) http://blog.ericgoldman.org/archives/2013/09/craigslist_wins_1.htm.

72. *Craigslist*, 942 F. Supp. 2d at 970.

73. *See* Facebook, Inc. v. Power Ventures, Inc., 844 F. Supp. 2d 1025, 1038–39 (N.D. Cal. 2012).

74. CFAA, 18 U.S.C. § 1030(c) (2012).

have a long duration. However, data hosts are unlikely to prevail on copyright claims if they do not control the copyright of all the works or if the works are not registered appropriately. Although scrapers may have valid fair use defenses,[75] this Section only discusses the challenges data hosts face in asserting valid copyright claims.

Copyright protection is widely applicable and covers a broad range of subject matter, but databases are often not protectable because of their unprotectable facts. Generally, "[c]opyright protection subsists . . . in original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device."[76] This requirement has been interpreted to require both fixation and originality.[77] Facts cannot be protected, as they are discovered and not created, but compilations of facts may receive protection if sufficiently original.[78]

In *Craigslist*, the court found that the user-generated ads on the Craigslist site were sufficiently original in their organization and that they could be copyrighted if properly registered.[79] Even though the ads contained factual information, the court found that the arrangement of this information in a context chosen by users was sufficiently different from raw factual information that could not be protected.[80] Courts have previously found classified advertisements eligible for copyright protection, though the copyright for an assortment of ads did not extend to individual ads.[81]

Courts in MLS cases have found various degrees of protections for different portions or reposted real estate ads.[82] The success of these suits

75. In the *Craigslist* case, there is a fair use argument that users who created ads can republish them on other websites, despite Craigslist's exclusive license to them. *See* Stephanie Marie Davies, *Rants and Raves: Craigslist's Attempt to Stop Innovating Third-Party Web Developers with Copyright Law*, 20 J. INTELL. PROP. L. 379 (2013); *see also* Kelly v. Arriba Soft Corp., 336 F.3d 811, 815 (9th Cir. 2003) (finding thumbnail photos were a transformative use and benefited from the fair use defense).

76. 17 U.S.C. § 102(a) (2012).

77. Fixation has been interpreted as requiring embodiment. *See* MAI Sys. Corp. v. Peak Computer, Inc., 991 F.2d 511, 518 (9th Cir. 1993).

78. The Copyright Act of 1976 specifically mentions compilations as being able to receive protection. *See* 17 U.S.C. § 103; Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., Inc., 499 U.S. 340, 345 (1991).

79. *See* Craigslist, Inc. v. 3Taps, Inc., 942 F. Supp. 2d 962 (N.D. Cal Apr. 30, 2013).

80. *Id.*

81. *See, e.g.*, Want Ad Digest, Inc. v. Display Adver., Inc., 653 F. Supp. 2d 171, 175 (N.D.N.Y. 2009).

82. *See* Jennings, *supra* note 15, at 141–42.

often depends upon the type of reposted information. For example, when the MLS reposts pictures, the MLS is more likely to infringe copyright.[83] The verbal description of real estate ads, however, is less certain since the property descriptions contain both factual and nonfactual material. Although the statements often contain non-copyrightable facts, they also contain artful descriptions that may be eligible for protection.[84]

Further increasing the burden on data hosts, the requirements for what constitutes a compilation work may be increasing. The U.S. Copyright Office published a new statement of policy restricting the scope of compilations.[85] It is also unclear if registration applications must include all individual authors in a compilation. For example, Craigslist registered their site as a compilation work, thus serving to register all component works to which Craigslist had an exclusive license.[86] The court looked to the Copyright Office's interpretation of the registration requirement and determined that "despite the omission of individual authors from the registration application" the registration of the collective work was still sufficient.[87] The court noted that given the volume of ads that Craigslist might have an exclusive license for, it might be "inefficient" to require registration of each individual author.[88] This logic may protect other data hosts who seek copyright protection for vast amounts of works.

Despite the Copyright Office's seeming willingness to allow the registration of large compilation copyrights, data hosts may struggle to achieve ownership of underlying data.[89] UGC generates unique copyright ownership problems for data hosts including public relations backlashes. Courts are split on whether a failure to adequately protect data constitutes an

83. *See, e.g.*, Metro. Reg'l Info. Sys., Inc. v. Am. Home Realty Network, Inc., 722 F.3d 591 (4th Cir. 2013).

84. *Id.*

85. Registration of Claims to Copyright, 77 Fed. Reg. 37,605, 37,606 (June 22, 2012) (to be codified at 37 C.F.R. pt. 201) ("[I]f a selection and arrangement of elements does not result in a compilation that is subject matter within one of the categories identified in section 102(a), the Copyright Office will refuse registration.").

86. The court found that registration of the ads through compilation copyright was sufficient for protection. *See* Craigslist, Inc. v. 3Taps, Inc., 942 F. Supp. 2d 962, 970–72 (N.D. Cal. 2013).

87. *Id.* at 972.

88. *Id.* at 974–76.

89. Electronic transfer of copyright is probably legal but may not be practical. To have sufficient legal standing to sue, a website must have an exclusive license to UGC, and gaining such a license may cause public relations issues. *See* Jeff Neuburger, *Staving Off Scrapers of User-Generated Content with Electronic Copyright Transfers . . . a Legal (But, Perhaps Not a Practical) Solution*, LEXOLOGY (Nov. 12, 2013), http://www.lexology.com/library/detail.aspx?g=8e782324-659e-43ee-aa8b-88f83cebd31f.

implied license for scrapers.[90] Generally, a data host must have an exclusive license from the content creator to be able to sue.[91] Transfer of copyright ownership requires a written agreement signed by the copyright owner.[92] Agreements can only be considered "a writing" to grant an exclusive license if the agreement demonstrated the parties' intent to transfer a copyright.[93] Provided the click-wrap or browse-wrap constitutes a written agreement, user acceptance of that agreement can suffice as an electronic signature.[94] Electronic copyright transfers can grant an exclusive license.[95]

For example, in the Craigslist litigation, Craigslist's TOU alone did not specifically grant an exclusive license to sue on behalf of the content creators.[96] Although a non-exclusive license would be insufficient to sue, Craigslist attempted to sue over a subset of ads created when users had also agreed to a click-through agreement specifically granting an exclusive license.[97] Data hosts can only sue over UGC when they have an exclusive license, and this requires carefully formatted TOUs and click-through agreements.

Yet data hosts have struggled to gain exclusive licenses to content without raising alarm from consumer groups and generating public relations backlashes. Major data hosts have faced large press-relations issues when

90. Websites can use the robots.txt protocol to tell scrapers they do not wish to be scraped. Although this protocol is voluntary—i.e., it can be ignored by crawlers—courts often treat the use or failure to use this protocol as legally significant. *See supra* notes 38–40 and accompanying text. Courts remain split if failure to use a robots.txt protocol, or poor implementation of such a protocol, constitutes an implied license for scrapers. *See* Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537, 563 (S.D.N.Y. 2013) (finding no implied license when no robots.txt protocol was used). *But see* Parker v. Yahoo!, Inc., No. 07-2757, 2008 WL 4410095 (E.D. Pa. Sept. 25, 2008) (finding that the failure to use a robots.txt protocol constituted an implied license for scrapers).

91. Only the valid owner of the copyright or the grantee of an exclusive license can sue for copyright infringement. *See* Davis v. Blige, 505 F.3d 90, 100 (2d Cir. 2007).

92. *See* 17 U.S.C. § 204(a) (2012).

93. *See id.*; Radio Television Espanola S.A. v. New World Entm't Ltd., 183 F.3d 922, 927 (9th Cir.1999) (internal citations omitted) ("Rather, the parties' intent as evidenced by the writing must demonstrate a transfer of the copyright.").

94. *See* Metro. Reg'l Info. Sys., Inc. v. Am. Home Realty Network, Inc., 722 F.3d 591 (4th Cir. 2013) (finding that user uploads of images sufficed to transfer an exclusive license for copyright).

95. The acceptability of electronic signatures depends upon state contract law. Intent to sign is required for click-wrap transactions to constitute electronic signatures. *See* UNIF. ELEC. TRANSACTIONS ACT § 9 (1999).

96. *See Terms of Use*, CRAIGSLIST, http://www.craigslist.org/about/terms.of.use (last visited Dec. 19, 2013).

97. *See* Craigslist, Inc. v. 3Taps, Inc., 942 F. Supp. 2d 962 (N.D. Cal Apr. 30, 2013).

changing their TOUs or privacy policies.[98] Craigslist quickly removed its exclusive license provision after user outcry.[99] Although an exclusive license is necessary for a data host to bring a copyright claim, it is unclear how to gain one without upsetting users.

In summary, data hosts' copyright claims are stronger than one would expect given cases like *Feist* and *Lotus*.[100] These claims require the underlying content to be sufficiently creative, and factual material is ineligible for protection. The line between factual material—which cannot be protected—and creative material—which can be—is occasionally murky and adds ambiguity to the protections a data host may receive. Further, UGC data is increasingly prevalent online and poses new copyright registration challenges to data hosts. Data hosts can require copyright ownership transfers of UGC through click-wraps, but doing so may generate a negative press backlash. Copyright protection is appealing, but these limitations have caused data hosts to turn to other protections as well.

3. Electronic Trespass to Chattel

Data hosts that are directly harmed by scrapers can bring trespass to chattel claims. These claims require the data host to prove the scraper has harmed them. The success of these claims in the Ninth Circuit may be decreasing after the recent *Intel v. Hamidi* ruling that server inconveniences did not constitute an actionable harm.[101] Sophisticated scrapers are unlikely to actually crash a data host's server, which makes trespass claims seem increasingly unlikely to succeed.

Electronic trespass claims are essentially property law trespass claims that have been ported to digital and electronic contexts. Electronic trespass to

98. *See* Jeremy C. Owens, *Biz Break: Google, Facebook, and Instagram Risk User Backlash with Privacy Changes*, SAN JOSE MERCURY NEWS (Oct. 11, 2013, 4:15 PM), http://www.mercurynews.com/60-second-business-break/ci_24292785/biz-break-google-facebook-and-instagram-risk-user; *see also* Julianne Pepitone, *Instagram Can Now Sell Your Photos for Ads*, CNN MONEY (December 18, 2012), http://money.cnn.com/2012/12/18/technology/social/instagram-sell-photos (describing user backlash over Instagram's TOU allowing sale of user photos for ads).

99. *See* Kurt Opsahl, *Good News: Craigslist Drops Exclusive License to Your Posts*, ELECTRONIC FRONTIER FOUND. (Aug. 9, 2012), https://www.eff.org/deeplinks/2012/08/good-news-craigslist-drops-exclusive-license-your-posts; *see also* Nathan Matisse, *Craigslist Backtracks, Drops Exclusive License on Posts*, ARS TECHNICA (Aug. 9, 2012, 2:51 PM), http://arstechnica.com/tech-policy/2012/08/craigslist-backtracks-drops-exclusive-licensing-on-posts.

100. *See supra* note 50 and accompanying text.

101. Intel Corp. v. Hamidi, 71 P.3d 296 (Cal. 2003) (holding that one employee sending disruptive emails to fellow employees does not constitute trespass).

chattel claims require a "tangible interference" that obstructs a possessory interest.[102] The current legal standard for electronic trespass to chattel claims was developed in *eBay v. Bidder's Edge*.[103] The court set out a two-part test for a trespass to chattel claim, which requires that "(1) [the] defendant intentionally and without authorization interfered with plaintiff's possessory interest in the computer system; and (2) [the] defendant's unauthorized use proximately resulted in damage to plaintiff."[104] Data hosts often attempt to show the scraper caused harm by demonstrating interference with servers or similar technical difficulties.[105]

The standard for what constitutes harm has recently been raised in the Ninth Circuit.[106] In *Intel Corp. v. Hamidi*, the court found that minor interference with server usage was not sufficient to constitute an actionable harm.[107] However, the court left open the possibility that a greater interference, perhaps crashing a website's server, may still be an actionable harm under trespass to chattel.[108] It is unclear how courts will interpret this new standard and what type of harm data hosts must prove.

In *Craigslist*, the court found that the scraper may have limited or diverted server resources from Craigslist's servers.[109] The district court is still in the early stages of litigation, and the diversion of server resources may be rejected like the "minor interference" with server resources that was found insufficient in *Hamidi*. The court construed the magnitude of the server harm as a factual question that would be determined later in the litigation, and this potential harm was thus sufficient to deny the motion to dismiss on the trespass to chattel claim.[110] While it remains unclear if Craigslist will prevail with a trespass claim on a server harm argument, the district court's seeming reluctance to dismiss the trespass claim may represent a lower bar for harm than the clear language from *Hamidi* would suggest.[111]

---

102. *See* Ticketmaster Corp. v. Tickets.com, Inc., No. CV99-7654, 2003 WL 21397701, at *5 (C.D. Cal. Mar. 7, 2003) (granting summary judgment dismissing trespass to chattels for a failure to show physical harm or an impairment of servers).

103. eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1069–70 (N.D. Cal. 2000).

104. *Id.*

105. *See* Snap-on Bus. Solutions Inc. v. O'Neil & Assocs., Inc., 708 F. Supp. 2d 669 (N.D. Ohio 2010) (denying summary judgment when plaintiff showed defendant had crashed their servers).

106. *See Hamidi*, 71 P.3d 296.

107. *Id.*

108. *Id.* at 304–05.

109. Craigslist, Inc. v. 3Taps, Inc., 942 F. Supp. 2d 962 (N.D. Cal. 2013).

110. *Id.* at 966.

111. The *Craigslist* court's reluctance to dismiss the seemingly weak electronic trespass claim may signal data hosts to bring further server harm arguments against scrapers in the

The continuing success of trespass to chattel claims is unclear. If determination of server harm is a factual question, it will be difficult for scrapers even to have weak claims dismissed at an early stage. Scrapers may not be willing or able to sustain a case to the point where trespass claims would be adjudicated. Yet data hosts will have an increasingly difficult time proving harm. Scraping technology becomes increasingly more sophisticated and servers more robust. Savvy scrapers can decrease their impact on target websites by decreasing the frequency of data pulls, using multiple IPs, and spreading the load across multiple servers.[112] Given the uncertainty of trespass claims, data hosts continue to rely upon contractual remedies to prevent data scraping.

D. DEFENSIVE MEASURES TAKEN BY DATA HOSTS

To compensate for a lack of explicit legal protections, data hosts use contractual and technological methods to control third-party access to their data. Some of these defensive measures can be implemented before any scraping has occurred. These may include password protection, required login, TOUs, and mandatory click-through agreements. Often a data host will implement all of these protections to give themselves the broadest legal protections. If a data host determines that scraping has occurred, the data host can then take additional steps to stop it.

Legally, TOUs and "click-through" agreements control how scrapers access a data host. TOUs are typically listed at the bottom of a webpage and passively display terms that users agree to by using a site or benefiting from its services.[113] Click-through agreements require active input from a user and often provide comprehensive restrictions to delimit the permissible scope of user activity.[114] Both the TOU and click-through agreement may even directly prohibit the gathering of data from the website. Although these provisions

future. The factual requirements set by the court in proving harm will undoubtedly have ramifications in future cases.

112. With more resilient servers designed to handle higher volume, scrapers are less likely to cause harm. Note that trespass requires an actual harm, so simply detecting a scraper will be insufficient. Scrapers often share advice on how to avoid interfering with servers, which could result in them being detected and then blocked. *See, e.g.*, *How to Crawl Websites Without Being Blocked*, WEBSCRAPING.COM (Feb. 8, 2010), http://webscraping.com/blog/How-to-crawl-websites-without-being-blocked.

113. For a sample terms of use, see *Terms of Use*, CREATIVE COMMONS, http://creativecommons.org/terms (last updated Dec. 5, 2013).

114. Apple's privacy policy describes the click-through agreement and restrictions imposed on users. *See Privacy Policy*, APPLE, http://www.apple.com/privacy (last updated Aug. 1, 2013).

generally restrict users, they also help protect against malicious uses of data that may harm user experiences.[115]

Click-through agreements formatted to give sufficient notice to users are generally found enforceable by courts.[116] But it is important for a data host to properly format the TOU for it to be binding.[117] A data host should implement a click-through agreement that requires an affirmative action by users instead of a "browse wrap" that users can passively ignore.[118] Generally, it is important to display the terms prominently and avoid terms granting unilateral amendment powers. Click-through agreements should require affirmative action by users that unambiguously signify assent.[119] Even without a mandatory click through, websites that clearly display their TOUs may deter scrapers simply by warning that the gathering of data is impermissible.

Data hosts also seek to detect when scraping is occurring. Scraping may slow the processing power of websites, directly impair access to a website, or, in rare circumstances, crash a website or server.[120] Although this heavy-handed scraping is less common, hosts can monitor the reposting of their website data to detect commercial competitors that use a website's data for their own commercial ends. Online guides teach data hosts how to track republication of specific kinds of data.[121] For example, scrapers use

115. Contractual violations of TOUs have been alleged against bots that harm user experiences. *See* MDY Indus., LLC v. Blizzard Entm't, Inc., 629 F.3d 928, 954 (9th Cir. 2010) (bringing suit against a bot that allowed for automated World of Warcraft gameplay); Ticketmaster L.L.C. v. RMG Technologies, Inc., 507 F. Supp. 2d 1096, 1102 (C.D. Cal. 2007) (bringing suit against a bot that allowed purchasing of tickets before other customers).

116. Courts are willing to enforce click-through agreements so long as they conform to basic contract law principals such as consideration and intent to be bound. *See* Register.com, Inc. v. Verio, Inc., 356 F.3d 393, 403 (2d Cir. 2004). A survey of many such cases finds no cases where click-through agreements were found unenforceable; it is even accepted by courts that users are unlikely to read click-wraps and that this does not limit their enforcement. *See* Mark A. Lemley, *Terms of Use*, 91 MINN. L. REV. 459 (2006).

117. In the *Zappos* case, the TOU was overbroad and reserved a full right to modify the TOU at any time. The court found this TOU ineffective. *See In re* Zappos.com, Inc., Customer Data Sec. Breach Litig., 893 F. Supp. 2d 1058 (D. Nev. 2012).

118. For a more thorough discussion of the mistakes Zappos made, as well as an overview of what data hosts should do to bind users, see Eric Goldman, *How Zappos' User Agreement Failed in Court and Left Zappos Legally Naked*, FORBES (Oct. 10, 2010, 12:52 PM), http://www.forbes.com/sites/ericgoldman/2012/10/10/how-zappos-user-agreement-failed-in-court-and-left-zappos-legally-naked.

119. *See id.*

120. *See* Angwin & Stecklow, *supra* note 11.

121. Bloggers that use Blogspot can use the host website data metrics to track scraping of their posts. This tracking can include easy-to-use Google alerts, RSS feeds, and other aggregation tools. User-friendly guides instruct bloggers how to prevent and even monetize

automated processes or bots that gather data at rates far faster than humans. These data usage spikes make it fairly easy for data hosts to determine which IP addresses are scraping data from the site.[122]

Once data hosts are aware that scraping is occurring, they can take simple steps to stop scrapers. Data hosts can send cease-and-desist letters to the scrapers and may restate portions of the TOU that have potentially been violated.[123] Cease-and-desist letters put the scraper on legal notice of any alleged violations.[124] The data hosts may also take a variety of technological measures to deny access, including blocking the IP addresses that are gathering the data.[125] Data hosts can then bring legal claims against scrapers. If a scraper is accessing content that contains underlying intellectual property rights, violates a data host's TOU, or harms the data host during acquisition, the scraper will likely lose in court.

Data hosts can increase control of their data by establishing prophylactic defensive measures against scraping, vigilantly detecting scrapers, and bringing legal and technological action when scraping has been detected. With these measures data hosts have been generally successful in actions against scrapers, though the success of legal claims are dependent upon the underlying IP protection in the data and proper implementation of defensive measures, such as TOUs and click-through agreements. Yet before bringing any of these claims, a data host should also consider the scraper's intended use for the data. As discussed in Part II, *infra*, the data host should only exercise these legal options if scrapers seek to challenge the data host's business model parasitically and not to augment it mutualistically.

scraping of their content. *See* Kristi Hines, *Content Scrapers—How to Find Out Who Is Stealing Your Content & What to Do About It*, KISSMETRICS, http://blog.kissmetrics.com/content-scrapers (last visited Jan. 30, 2014).

122. Because bots gather data faster than humans, detection of scraping normally looks at the time interval between data requests from given IP address. These settings can be configured to change the interval timing and to specify a threshold that flags a potential bot or scraper. *See* IMPERVA, DETECTING AND BLOCKING SITE SCRAPING ATTACKS (2011), http://www.imperva.com/docs/wp_detecting_and_blocking_site_scraping_attacks.pdf; *see also* John Wagnon, *More Web Scraping—Bot Detection*, DEVCENTRAL (May 22, 2013), https://devcentral.f5.com/articles/more-web-scraping-bot-detection#.UrPFL2RDt10.

123. *See* Craigslist, Inc. v. 3Taps, Inc., 942 F. Supp. 2d 962, 966–67 (N.D. Cal Apr. 30, 2013).

124. Goldman, *supra* note 71.

125. Simple Java or PHP code can be used to block target IP addresses. Once again, web guides provide step-by-step instructions that can be followed with a modicum of programming knowledge. *See, e.g.*, *System: Blocking Unwanted Spiders and Scrapers*, THE ART OF WEB, http://www.the-art-of-web.com/system/block-spiders (last visited Jan. 30, 2014).

II. FACTORS SUPPORTING DATA HOST ENFORCEMENT

Even in circumstances where data hosts might prevail against scrapers, a nuanced study of the benefits of scraping within the community suggests that scrapers are increasingly mutualistic, adding value to data hosts and users. To maximize potential business opportunities, data hosts should be cognizant of user demands, potential benefits of scrapers, and public relations implications before bringing suits. If the scraper intends to create a service that augments the data host's services, then bringing a suit against the scraper may be shortsighted. If, however, the data host determines that the scraper poses a severe business threat and seeks to supplant or abuse the data host's services, legal action may be warranted.

When scrapers seek to replace or replicate the service offered by the data hosts, the data hosts should consider suing the scraper. Early in the history of online auction sites, a website called Bidder's Edge sought to aggregate online auction data from online auction sites including eBay.[126] Although users still had to proceed to eBay to buy the items, eBay saw this as a threat to their core business model. And although Bidder's Edge represented only a small percentage of total traffic, the court found that allowing Bidder's Edge to scrape would lead to a slippery slope where others would scrape eBay as well.[127] Despite aggregating other auction sites, Bidder's Edge represented not an augmentation of eBay's services, but potentially a direct threat. With little competition, eBay was able to become the dominant online auction site. If eBay had not sued, it is possible Bidder's Edge could have offered its own auction-hosting capabilities in the future to challenge eBay.

Parasitic threats from scrapers have also come in the form of challenges to online price discrimination and information control models. Epitomizing the classic dichotomy of "information wants to be free . . . information wants to be expensive," brokerage houses offering exclusive information to preferred customers have struggled to control widespread release of that information.[128] For example, financial brokerage houses would release early stock tips to preferred customers to create an incentive for using that firm to purchase stocks. Responding to high demand for this information, scraping services began to disseminate those stock tips to outside investors, stymying efforts to preserve the price discrimination model of the brokerage houses.

126. The court found that Bidder's Edge pulling of data from eBay nearly one hundred thousand times per day could constitute a trespass. *See* eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1060 (N.D. Cal. 2000).

127. *Id.*

128. STUART BRAND, THE MEDIA LAB 202, 211 (1987).

Even when just one scraper reposted the stock tips, the information lost all of its value. The brokerage houses unsuccessfully sought legal recourse against the data scrapers using misappropriation (and copyright) doctrines.[129] Furthermore, factual stock tips cannot be easily protected under IP law. The brokerage house must absolutely control their data, or else the preferred information model fails completely. This seems to be a losing proposition: new scraping services can spring up before brokerage houses can shut them all down. Data hosts in this circumstance are unable to prevent parasitic scraping if the incentives attracting new scrapers are too high.

Free rider problems continue to motivate some data hosts to sue, even in the face of overwhelming user demand for cheaper or better services. Nevertheless, to determine the proper amount of permissible scraping, free riding concerns must be weighed against the public benefits of scraping. For example, in online real estate databases, MLS real estate sites continue to resist data-scraping services. MLS services aggregate real estate listing information, allow access to that data for a fee, and tend to drive up competition in rental and housing markets by increasing user information.[130] MLS sites sue scrapers, claiming the reposting of housing descriptions, pictures, or listing information violates the legal protections that exist in those elements.[131] Data hosts—the MLS services—seek to prevent the scrapers from free riding and reposting the information that they charge for others to access. MLS sites argue scrapers are essentially poaching their work without paying for it. Unfortunately for MLS sites, users want free aggregated listings.[132] Like brokerage houses, MLS sites may win in court, but user demand will spur the growth of new free MLS sites to replace the old. It is clear that user demand for increased transparency and superior services will continue to drive consumer demand for free MLS services, and many suits by MLS services are still pending.[133]

---

129. Brokerage houses failed to protect their data with hot news misappropriation theories. The court found misappropriation required passing off content as one's own, which Fly was not doing. Allowing scraping contingent upon proper attribution also represented a shift away from a commercial assessment of any "free riding." *See* Barclays Capital Inc. v. Theflyonthewall.com, Inc., 650 F.3d 876 (2d Cir. 2011); *see also* Anthony Corleto & Scott Smedresman, *Second Circuit Decision in Barclays v. Theflyonthewall.com Sheds Light on Conflict over "Hot News,"* BLOOMBERG L. (SEPT. 13, 2011), http://about.bloomberglaw.com/practitioner-contributions/second-circuit-decision-in.

130. *See* Jennings, *supra* note 15, at 1–3.

131. *Id.*

132. *Id.*

133. MLS sites continue to take action against data scrapers. Their efforts include identifying and blocking scrapers, as well as exercising IP rights against them. *See Industry*

Scraping is often beneficial, and data hosts must realize that user demand may drive change, even when faced with resistance from data hosts. The user value offered by financial aggregators represents just one beneficial application of scraping. At the turn of the century, many large financial institutions were increasingly resistant to the efforts of financial aggregators.[134] These aggregators created websites, and now apps, that allowed users to view all their financial data in one place instead of having to log into each of their banking, billing, or checking accounts.[135] Financial institutions aggressively sought to stop the spread of financial aggregators; they feared that, if customers could access their data outside of banking websites, customers would be less likely to use online banking products that benefitted the banks themselves.[136] Despite this resistance, user demand for efficient and secure financial aggregation was overwhelming.[137] Customer demand eventually won out and now most major banks offer their own financial aggregation services. Some financial institutions now directly partner with aggregators to provide users with the best possible services. Initial resistance to aggregation has now turned to fierce competition between aggregation services with each bank seeking to provide the best scraping services. Banks are still able to offer online services, and users now benefit from the convenience of having their financial information securely aggregated thanks to partnerships between scrapers and data hosts.[138] Instead of fighting the development of financial aggregators, data hosts who partnered with the scrapers early on could have gotten ahead of this trend and even attracted new users.

Like data hosts who successfully paired with financial aggregators, data hosts should embrace mutualistic scrapers that seek to improve their underlying services. Internet travel purchased through airline aggregators demonstrates that when data hosts tolerate augmentation by scraping, both users and the data hosts stand to benefit. Before airline aggregators, users were forced to individually compare travel options at each airline's website.[139] Airline aggregators now scrape data from carrier websites to offer price

Cracks Down on Listing Scraping, REALTOR MAG (Aug. 1, 2013), http://realtormag.realtor.org/daily-news/2013/08/02/industry-cracks-down-listing-scraping.

134. Sult, *supra* note 16, at 20.

135. Julia Alpert Gladstone, *Data Mines and Battlefields: Looking at Financial Aggregators to Understand the Legal Boundaries and Ownership Rights in the Use of Personal Data*, 19 J. MARSHALL J. COMPUTER & INFO. L. 313 (2001).

136. *Id.*

137. *Id.*

138. A customer survey indicated that security was a key concern when choosing financial aggregators. *See id.*

139. *See* Phil Cameron, *Internet Travel Purchases*, GPSOLO, May/June 2013, at 48.

comparisons among different travel options,[140] allowing users to find and purchase the cheapest flights and bundle rental cars, hotels, and flights.[141] The data hosts, here the airlines that have the fare information, also benefit through increased visibility, advertising, and search engine optimization ("SEO") that aggregators perform.[142] Importantly, the scrapers are attempting to add value to the airline's offerings, not to supplant them directly. Aggregators do not actually seek to fly travelers around; they merely make travel purchases easier, earning money from referral fees.[143] This mutualism benefits both the data hosts—the air carriers—and the scrapers—the airline aggregators.

Yet, despite the apparent value to users, not all airlines have warmed to interacting with scrapers. While larger airlines have generally embraced airline aggregation, some budget carriers have excluded themselves from airline aggregators. These are often budget airlines that seek to keep costs as low as possible.[144] By removing themselves from aggregators, airlines can avoid the

140. Although there are many airline aggregators, as of 2013 Expedia is the largest. *Expedia, Inc. Overview*, EXPEDIA INC., http://www.expediainc.com/about/ (last visited Mar. 5, 2014). Most operate similarly and many aggregators have paired with each other to offer more comprehensive search options to users. *See About Expedia.com*, EXPEDIA, http://mediaroom.expedia.com/about-expediacom-180 (last visited Jan. 30, 2014); *see also* ARTHUR FROMMER, ASK ARTHUR FROMMER: AND TRAVEL BETTER, CHEAPER, SMARTER 1–14 (2009) (passing on detailed knowledge of airline aggregators and pricing models directly to customers as a way to heavily defray the costs of travel); Ed Hewitt, *The Aggregators Are Coming*, INDEPENDENT TRAVELER, http://www.independenttraveler.com/travel-tips/travelers-ed/the-aggregators-are-coming (last visited Jan. 30, 2014).

141. Many travel guides now recommend that customers visit aggregators before visiting individual airlines websites. Indeed guides may caution that airlines not listed by aggregators tend to be smaller, limited in scope, or regional. *See, e.g.*, Pete Werner, *How to Beat the High Cost of Airfare*, WALT DISNEY WORLD INFO, http://www.wdwinfo.com/discounts/beat-high-cost-airfare.htm (last visited Jan. 30, 2014); *see also* Ben Mutzabaugh, *A Guide to Booking Air Tickets Online*, USA TODAY (Oct. 23, 2003, 1:32 PM), http://usatoday30.usatoday.com/travel/tips/booking/2003-10-23-airline_x.htm.

142. A travel industry study examined the social media popularity and search engine presence of common aggregators and air travel providers. The study concluded that it was more efficient for aggregators to invest in SEO and many forms of online advertising than air travel providers. Aggregators had the higher incentives to invest in online ads as they seek to attract a broad range of customers, while air travel providers are typically limited by geography and scope and do not benefit from widespread advertising. *See Travel Industry Study*, SEARCHMETRICS (Nov. 2, 2013), *available at* http://www.searchmetrics.com/en/white-paper/travel-industry-flights.

143. Airline aggregators make money when users click through their site and purchase a ticket. The costs of these referral fees are often passed on to the airlines as a cost of being included in the aggregation. *See* FROMMER, *supra* note 140, at 1–4.

144. Budget carriers, like Southwest, JetBlue, and Ryanair, often have a rocky relationship with aggregators. Aggregators often charge fees that the budget airlines object to, and many budget airlines try and drive revenue solely through their websites, though that

small click-through referral fee charged by the Global Distribution Services that sell fare information to aggregators.[145] Airlines have been successful in removing themselves from aggregators through legal action.[146] Airlines that are not displayed in aggregators do not benefit from the advertising of aggregators, but have more control over the display of their information. Yet trying to drive all ticket sales through a carrier's website can draw the ire of customers and consumer groups.[147] Larger airlines have now largely accepted aggregators, and many contract directly with aggregators to ensure they are offering the best services to their users.[148] The budget carriers are betting that users would rather have marginally cheaper fares instead of the convenience of aggregation. This business decision requires users to make the same value calculation as the airlines: the users must be lured by cheaper prices to take the extra time to visit budget carrier's sites individually. Data hosts weigh the mutualistic benefits of working with scrapers against those costs.[149]

Data hosts should also be careful when suing scrapers that are widely popular. Public support can be a decisive factor when deciding to sue; even if data hosts shut down scrapers, new ones will spring up to take their places. This has largely been the case in Craigslist's litigation battles. In *Craigslist*, users directly benefited from the 3Taps and Padmapper map functions.

position may be softening. *See* Kevin May, *Ryanair Offers Meta Search Engines an Olive Branch*, TRAVOLUTION (Sept. 2, 2008, 10:48 AM), http://www.travolution.co.uk/articles/2008/09/02/1680/ryanair-offers-meta-search-engines-an-olive-branch.html.

145. Tom Lee, *The Airline's Ongoing Struggle with Price Aggregation Sites*, TECHDIRT (July 29, 2008, 1:50 AM), http://www.techdirt.com/articles/20080725/1322411794.shtml.

146. Southwest has brought suits to be excluded from airline aggregators. *See* James Peltz, *Southwest Airlines Sues Orbitz Travel Site*, LOS ANGELES TIMES (May 5, 2001), http://articles.latimes.com/2001/may/05/business/fi-59592; *see also* Jennifer Disabitino, *Southwest Sues Orbitz over Flight Information*, CNN (May 8, 2001, 11:34 PM), http://edition.cnn.com/2001/TECH/industry/05/08/southwest.sues.orbitz.idg.

147. Ryanair, a budget European carrier, threatened to remove its fares from all third-party sellers and even threatened to rescind tickets that had been purchased through those sellers. *See* Laura Noonan, *Ryanair Travelers May Lose Bookings*, IRISH INDEPENDENT (Aug. 8, 2008), http://www.independent.ie/business/irish/ryanair-travellers-may-lose-bookings-26467772.html.

148. *See* Christopher Hinton, *US Airways Inks Contract with Expedia*, WALL ST. J.: MARKETWATCH (Jan. 21, 2011, 3:16 PM), http://www.marketwatch.com/story/us-airways-inks-contract-with-expedia-2011-01-21 (describing the costs to airlines from third-party scrapers).

149. In another example of a data host weighing the benefits of mutualistic scraping, Apple shut down scraping services that helped users find availability of Apple products at Apple stores. *See* Don Reisinger, *Two Tracker Services Shutter, Following Apple's DMCA Takedown*, CNET (Nov. 6, 2013, 7:45 PM), http://news.cnet.com/8301-13579_3-57611069-37/two-tracker-services-shutter-following-apples-dmca-takedown. Although this undoubtedly helped customers find the products they were looking for, it may have taken traffic away from Apple's site or raised other concerns. *See id.*

Regardless of the outcome, Craigslist has already been the target of public criticism for merely bringing the suits.[150] Although Craigslist is currently prominent in the online classifieds market,[151] sufficient public backlash increases the likelihood that a competitor could gain traction. User demand for new services should make data hosts more receptive to efforts from scrapers. The public support for 3Taps and PadMapper reflects user demand for a map service.[152] User demand for new features manifested years ago and could be easily measured by the popularity of Oodle, Listpic, and Flippity.[153] Yet instead of working with these third-party services, Craigslist mounted a series of expensive lawsuits against them. When faced with a scraping service that enjoys popular support, data hosts should consider more cooperative methods of dealing with scrapers.

Instead of resisting user demand and fighting trends with legal action, data hosts should look to adapt their business models and benefit from scrapers. Scraping typically occurs because there is no easy, or legal, way to access a data host's data. By setting up APIs, data hosts can encourage cooperative scraping and easily control and monitor access to their data.[154] Many large tech companies use APIs to better control third parties who might otherwise scrape their data.[155] APIs may also increase the perception that data hosts are innovative and open to third-party support. If a scraper is problematically scraping data from within an API, it is a simple matter for the data host to revoke access. There are now programming tools that decrease the technological hurdles of using APIs instead of scraping.[156] Data

150. *See* Mike Masnick, *Disappointing: Craigslist Sues Padmapper for Making Craigslist More Useful & Valuable*, TECHDIRT (Jul. 25, 2012, 7:12 AM), http://www.techdirt.com/articles/20120724/18071219816/disappointing-craigslist-sues-padmapper-making-craigslist-more-useful-valuable.shtml.

151. *Craigslist Factsheet*, *supra* note 20.

152. *See* Nick Bilton, *Disruptions: Innovations Snuffed Out by Craigslist*, N.Y. TIMES: BITS (July 29, 2012, 11:00 AM), http://bits.blogs.nytimes.com/2012/07/29/when-craigslist-blocks-innovations-disruptions.

153. ListPic, Oodle, and Flippity all received high web traffic before being shutdown. *See supra* Section I.A.

154. The Google AdWords API encourages developers to use AdWords data creatively, but does so in a way that Google can easily control and monitor. *See AdWords API*, GOOGLE, https://developers.google.com/adwords/api/index (last visited Jan. 31, 2014).

155. Google has exerted pressure on those who scrape AdWord data to use the AdWords API instead. *See* Patrick Keeble, *A Message from Our CEO: Raven to Drop Rankings, Other Scraped Data on Jan. 2*, RAVEN (Dec. 7, 2012), http://raventools.com/blog/scraped-data-serp-tracker; *see also* Sean Smith, *Google Declining API Access of AdWords to Data Scraping Services: No Need to Panic*, @SNSMTH (Dec. 7, 2012), http://www.snsmth.com/blog/google-declining-adwords-api-access-to-data-scraping-services.

156. The popular code instruction website Codecademy offers specific courses on how to work with popular APIs. *See* Ben Popper, *Codecademy Teaches Users to Work With APIs From*

hosts may finally be recognizing the benefits of cooperative scraping despite resistance from older data hosts.[157] Government organizations are also open to the benefits of API implementation.[158] Increased use of APIs by established corporations and government groups is a positive indicator that data hosts are increasingly open to mutualistic relationships with scrapers.

Data hosts should consider mutualistic relationships with scrapers before taking legal action. If properly protected by mandatory click-through agreements and TOUs, data hosts may win against scrapers in court despite public support for scrapers, user demand for new services, or competitive pressures. Yet these factors increase the likelihood that such suits will be ineffective in deterring new scrapers and may upset existing users. Further, an overly litigious data host may fall victim to competitive pressures from other services that are more willing to engage in mutualistic relationships with data scrapers that benefit users. Instead, data hosts should emphasize cooperative scraping and try and encourage services that are beneficial to both users and the data host. Data hosts can benefit from enhanced services and user support by channeling third-party developers and scrapers through APIs.

III. CONCLUSION

Currently, data hosts can mount strong legal defenses to scraping. By carefully controlling scraping in their TOUs, practicing IP-blocking, and sending cease-and-desist letters to scrapers, data hosts can bring strong CFAA claims against scrapers. If the scraping is poorly executed and crashes or impairs the data host's website, the data host will likely have a valid electronic trespass to chattel claim. If the data host owns intellectual property rights in the data they host, the data host can bring compilation copyright claims as well. If the scraper has violated a contractual obligation with the data host, the data host can also proceed on contract claims. These myriad legal claims can overwhelm the legal resources of scrapers and quickly drive them out of business.

YouTube, NPR, and More, THE VERGE (Jan. 9, 2013, 12:00 PM), http://www.theverge.com/2013/1/9/3855090/codecademy-teaches-users-to-work-with-apis-from-youtube-npr-and-more.

157. *See* Dan Woods, *Explaining the API Revolution to Your CEO*, FORBES (Dec. 15, 2011, 6:59 PM), http://www.forbes.com/sites/danwoods/2011/12/15/explaining-the-api-revolution-to-your-ceo. *See generally* DAN JACOBSON, APIS: A STRATEGY GUIDE (2011).

158. *See Benefit of APIs*, HOWTO.GOV, http://www.howto.gov/mobile/apis-in-government/benefits-of-apis (last visited Jan. 31, 2014).

Yet the legal doctrines involved in scraping suits are in flux. The currently strong claims brought by data hosts may not be on such firm ground in the near future. The CFAA, electronic trespass to chattels, and compilation copyright causes of action have all received recent attention and may be changing doctrinally. Perceived abuse or misuse of the CFAA has led policy makers and lawyers to consider curtailing its use.[159] The requirements for proving an actionable harm for electronic trespass to chattel has increased, and savvy scrapers have learned not to crash websites while scraping.[160] Further, getting users to agree to an exclusive license, which may be required for compilation copyright claims, has generated public outcry against typically sympathetic technology companies.[161] It remains unclear if these claims will continue to be as successful, or popular, with data hosts in the future.

Even though data hosts may prevail on legal grounds against scrapers, it may not be in their business interest to simply sue. Suing all scrapers fails to recognize the value that a mutualistic relationship with scrapers can have for a data host; the most business savvy data hosts should seek to guide scrapers to the most beneficial channels possible. Many industries that were initially resistant to scraping now benefit from a mutualistic relationship with them.[162]

Data hosts that accept the valuable role of scrapers in the digital environment stand to benefit from cooperative scraping. By offering API access to their data, data hosts can more easily monitor the scraping that does occur and guide the direction that scraping takes.[163] Further, promising uses of scraping can be brought in-house if the data host chooses to partner with the scraping service to open up new business models. With the rise of more

159. The CFAA has been prominently featured in two recent tragedies: the tragic suicide of a teen girl following Myspace teasing and the death of technology and privacy activist Aaron Swartz. Prosecutors alleged CFAA claims in both cases, and the use of an anti-hacking law as a prosecutorial tool has drawn substantial public criticism. *See* Jennifer Steinhauer, *Woman Found Guilty in Web Fraud Tied to Suicide*, N.Y. TIMES, Nov. 27, 2008, at A25; Tim Wu, *Fixing the Worst Law in Technology*, NEW YORKER (Mar. 18, 2013), http://www.newyorker.com/online/blogs/newsdesk/2013/03/fixing-the-worst-law-in-technology-aaron-swartz-and-the-computer-fraud-and-abuse-act.html; *see also US v. Drew*, ELECTRONIC FRONTIER FOUND., https://www.eff.org/cases/united-states-v-drew (last visited Jan. 31, 2014).

160. The higher standard for an actionable harm for electronic trespass to chattel increases the burden of the data host. *See* Intel Corp. v. Hamidi, 71 P.3d 296 (Cal. 2003).

161. *See* Masnick, *supra* note 59.

162. Airline and financial aggregators demonstrate potential benefits of working with scraping. *See supra* Part II.

163. *See* Manfred Bortenschlager, *Leveraging APIs as Part of Digital Strategy*, WIRED (Dec. 17, 2013), http://www.wired.com/insights/2013/12/leveraging-apis-part-digital-strategy; *see also Benefit of APIs*, *supra* note 158.

advanced data analytics, data hosts stand to benefit from scraping more so than ever before.[164]

Scrapers must still carefully examine their own business models. If a data host perceives a scraper as parasitic, the data host can deny the scraper access to their data and proceed with legal action. Although the data host may face public backlash for discontinuing a popular scraping service, scrapers should focus on adding value for users in a context that does not seek directly to undermine the data host. New data analytics and cooperation between data hosts and third parties offer new possibilities for mutualistic scraping that stands to benefit users, scrapers, and data hosts.

164. Cases have not considered scrapers that are using data to perform complex analytics. It remains unclear how legal regimes will adapt to uses of data that are truly novel. *See* Jim Snell & Derek Care, *Use of Online Data in the Big Data Era: Legal Issues Raised by the Use of Web Crawling and Scraping Tools For Analytics Purposes*, BLOOMBERG L., http://about.bloomberglaw.com/practitioner-contributions/legal-issues-raised-by-the-use-of-web-crawling-and-scraping-tools-for-analytics-purposes (last visited Feb. 13, 2014).