ARTICLE

TWENTY YEARS OF WEB SCRAPING AND THE COMPUTER FRAUD AND ABUSE ACT

ANDREW SELLARS†

I. INTRODUCTION

At the start of 2017, scientists worried that the incoming presidential administration would remove politically inconvenient environmental information from government websites.[1] In response a group of faculty and students formed the Environmental Data & Governance Initiative.[2] The initiative launched a number of projects to monitor and preserve federally-hosted scientific data on government databases.[3] This effort led the group to discover that the National Park Service had removed ninety-two documents related to efforts to reduce carbon emissions under the "Climate Friendly Parks Program," which (once confronted) the Service then promised to re-post.[4]

† Lecturer and Clinical Instructor, Boston University School of Law; Director, BU/MIT Technology & Cyberlaw Law Clinic. Thanks to Stacey Dogan, Jonathan Frankle, and Ahmed Ghappour for their valuable comments; to my symposium co-panelists Jamie Lee Williams and David Thaw, and moderator Paul Gugliuzza, for their thought-provoking contributions; and to current and former BU Law students Danielle Deluty, Kaitlin Heinen, Cliff Sonkin, and Yajing Wen for their research assistance.

[1] This concern bore out, Oliver Milman & Sam Morris, *Trump is Deleting Climate Change, One Site at a Time*, GUARDIAN (May 14, 2017) https://www.theguardian.com/us-news/2017/may/14/donald-trump-climate-change-mentions-government-websites [https://perma.cc/2P7C-YZA8], although the EPA did preserve a snapshot of the website as it existed on the last day of the prior administration, *see EPA's January 19 Snapshot*, EPA, https://19january2017snapshot.epa.gov/ (last visited Mar. 8, 2018) [https://perma.cc/G94G-GB9E].

[2] *About*, ENVTL. DATA & GOVERNANCE INITIATIVE, https://envirodatagov.org/about/ (last visited Mar. 6, 2018) [https://perma.cc/3KQ9-32SH].

[3] *Website Monitoring*, ENVTL. DATA & GOVERNANCE INITIATIVE, https://envirodatagov.org/website-monitoring/ (last visited Mar. 6, 2018) [https://perma.cc/D7TV-5MRL].

[4] Sarah Emerson, *The National Park Service Promises to Reinstate 92 Climate Change Documents Removed from Website*, MOTHERBOARD (Dec. 22, 2017), https://motherboard.vice.com/en_us/article/j5vpak/the-national-park-service-promises-to-reinstate-92-climate-change-pages-removed-from-website [https://perma.cc/E22X-DXYD].

A few years earlier, Alexis Madrigal sought to understand the strange "microgenres" that the online video platform Netflix creates for its users, like "Critically-Acclaimed Crime Movies from the 1940s," "Visually Striking Latin American Comedies," or "Suspenseful TV Shows Featuring a Strong Female Lead."[5] He discovered that Netflix hosted these genres on a section of its website, and by using a tool that gathered all 76,897 of them, Madrigal was able to learn how these classifications were structured and deployed.[6] He then used this research to secure an interview with the team at Netflix who developed the elaborate tagging system, and wrote up an extensive analysis of it for *The Atlantic*.[7]

Historian Jason Scott focuses instead on Internet preservation.[8] Scott is the founder of an online collective known as Archive Team, who watch for the closure of famous social websites from earlier days in Internet history and make as full a backup of the website as possible before the site is taken down.[9] Once-popular websites like Geocities, Friendster, and Miiverse, now gone from their original domains, are preserved by Archive Team, providing new opportunities for scholars to analyze these now-defunct platforms.[10]

Each of these projects was possible thanks to an Internet research technique known as "web scraping." Web scraping generally refers to the retrieval of content posted on the World Wide Web through the use of a program other than a web browser or an application programming interface (API).[11] In most cases this is done through a computer script that will send tailored queries to websites to retrieve specific pieces of content. These requests are often sent in an automatically generated series of requests, in order to extract material

[5] Madrigal maintains a public list of these "microgenres." Alexis Madrigal, *Netflix Microgenres*, GOOGLE DOCS, https://docs.google.com/spreadsheets/d/1XyswDlnyP6dlLyL8brhTFfTORFm-GZ2MnwTPEgp0HCc/ [https://perma.cc/VK5J-STBS] (last visited Mar. 6, 2018).

[6] Alexis C. Madrigal, *How Netflix Reverse Engineered Hollywood*, ATLANTIC (Jan. 2, 2014), https://www.theatlantic.com/technology/archive/2014/01/how-netflix-reverse-engineered-hollywood/282679/ [https://perma.cc/FW8W-XCNN].

[7] *Id*.

[8] *See* ARCHIVE TEAM, https://archiveteam.org/ (last updated June 28, 2015) [https://perma.cc/MS24-9JF7].

[9] Matt Schwartz, *Fire in the Library*, MIT TECH. REV. (Dec. 20, 2011), https://www.technologyreview.com/s/426434/fire-in-the-library/ [https://perma.cc/N3M4-HJEV].

[10] *See* Ian Milligan, *Finding Community in the Ruins of Geocities: Distantly Reading a Web Archive* (Oct. 2015), https://uwspace.uwaterloo.ca/bitstream/handle/10012/11650/milligan-s.pdf [https://perma.cc/RQ3U-VTLF].

[11] RYAN MITCHELL, WEB SCRAPING WITH PYTHON viii–ix (2015). Though as noted below, it goes by several other names as well, with disagreement as to whether the terms refer to different acts, and the precise definition can be somewhat fluid. *See infra* notes 57–82 and accompanying text.

across an array of websites or a large collection of material from a specific website.[12]

The technique has countless applications. It can be used to preserve websites,[13] help identify and extract data for analysis,[14] aggregate information from disparate sources,[15] and map out unexplored networks of servers and websites.[16] Its use can help competition by lowering startup information barriers,[17] enable consumers to find deals and discounts in online services,[18] identify and correct issues of algorithmic bias,[19] and introduce new forms of humor and playfulness.[20] (The technique is capable of less appealing uses as well. It can facilitate an invasion of one's sense of privacy,[21] expose content that a website host wished instead to remain hidden,[22] facilitate copyright infringement at

[12] *See infra* notes 101-116 and accompanying text.

[13] *See, e.g.*, PERMA.CC, https://perma.cc/ [https://perma.cc/MT6S-EQ3M] (last visited July 28, 2018) (providing a tool to archive websites in scholarly and judicial publications to avoid "link rot").

[14] *See generally* DANIEL T. LAROSE & CHANTAL D. LAROSE, DISCOVERING KNOWLEDGE IN DATA: AN INTRODUCTION TO DATA MINING (2d ed. 2014).

[15] KIMBERLY ISBELL & CITIZEN MEDIA LAW PROJECT, THE RISE OF THE NEWS AGGREGATOR: LEGAL IMPLICATIONS AND BEST PRACTICES 1–2 (2010), https://cyber.harvard.edu/publications/2010/news_aggregator_legal_implications_best_practices [https://perma.cc/7FSN-TFW7].

[16] *See, e.g.*, Qinqhua Zheng et al., *Learning to Crawl Deep Web*, 38 INFO. SYS. 801, 801 (2013).

[17] *See* Rory Van Loo, *Rise of the Digital Regulator*, 66 DUKE L.J. 1267, 1285–89 (2017).

[18] *See* Complaint, Sw. Airlines Co. v. Roundpipe LLC, No. 3:18-CV-33 (N.D. Tex. filed Jan. 5, 2018) (filed by an airline against a scraper that allowed customers to take advantage of rebooking deals within their reservation system).

[19] Amanda Levendowski, *How Copyright Law Can Fix AI's Implicit Bias Problem*, 93 WASH. L. REV. (forthcoming 2018) (arguing that broader access to datasets can help correct bias in how algorithms are currently trained).

[20] *See, e.g.*, *Chez LA*, COMEDY HACK DAY (Aug. 27, 2015), http://www.comedyhackday.org/demosmade/2015/8/27/chez-la [https://perma.cc/9FHJ-MN27] (aggregating information that aspiring young actors purportedly need for their move to Los Angeles — and then home again when dreams of stardom do not pan out); Erowid Coin Bot (@icowid), TWITTER, https://twitter.com/icowid [https://perma.cc/U862-YLPU] (last updated Mar. 2, 2018) (posting content scraped from whitepapers for cryptocurrency initial coin offerings, mixed with posts about of bad experiences from a recreational drug website, to highlight the cultish tendencies of both communities).

[21] Zachary Gold & Mark Latonero, *Robots Welcome? Ethical and Legal Considerations for Web Crawling and Scraping*, 13 WASH. J. L. TECH. & ARTS 275, 282–83 (2018); Joseph Cox, *70,000 OkCupid Users Just Had Their Data Published*, MOTHERBOARD (May 12, 2016), https://motherboard.vice.com/en_us/article/8q88nx/70000-okcupid-users-just-had-their-data-published [https://perma.cc/HZ2A-ESWT].

[22] *See, e.g.*, In re Complaint of Judicial Misconduct, 575 F.3d 279, 291 (3d Cir. 2009) (admonishing a judge for misconduct after he hosted pornographic images on a private web server, which were unintentionally indexed by search engines).

scale,[23] enable new forms of surveillance,[24] or help people cheat in online trivia games.[25]) Given its utility, the technique has been adopted widely. One company estimates that about a quarter of all current web traffic comes from web scrapers.[26]

Web scraping has proliferated beneath the shadow of the federal anti-hacking statute, the Computer Fraud and Abuse Act (CFAA).[27] For those who do not want their websites scraped, the CFAA presents a possible remedy through its broad prohibition against obtaining information by accessing a computer without authorization or by exceeding one's authorized access.[28] While first drawn to regulate only "federal interest computers,"[29] the statute grew to govern most Internet-connected computers by the late 1990s, when courts considered its application to web scraping.[30] Criminal cases have been brought against scrapers,[31] but the real area of growth has been with the

[23] Bob Bardwell, *Don't Get Scraped: Putting an End to Web Scraping, Content Theft*, RACKSPACE BLOG (June 14, 2012), https://blog.rackspace.com/dont-get-scraped-putting-an-end-to-web-scraping-content-theft [https://perma.cc/8VDZ-Q2SC].

[24] Jonathan Frankle, *How Russia's New Facial Recognition App Could End Anonymity*, ATLANTIC (May 23, 2016), https://www.theatlantic.com/technology/archive/2016/05/find-face/483962/ [https://perma.cc/D84K-GSLY].

[25] Aaron Mak, *Developers are Creating Bots that Can Help People Cheat at HQ Trivia*, SLATE (Jan. 24, 2018), https://slate.com/technology/2018/01/bots-can-greatly-assist-players-in-the-popular-hq-trivia-game.html [https://perma.cc/82F6-N9BK].

[26] The company found a little more than half of all web traffic as coming from bots, and a little more than half of those were what the company called "bad bots," which appear to be oriented toward taking down websites by overwhelming them with requests rather than retrieving information in a systemic way. Igal Zeifman, *Bot Traffic Report 2016*, IMPERVA INCAPSULA (Jan. 24, 2017), https://www.incapsula.com/blog/bot-traffic-report-2016.html [https://perma.cc/76NC-B4BV].

[27] Enacted first in 1984 under the Comprehensive Crime Control Act, and then expanded as the CFAA in 1986. *See* Orin Kerr, *Vagueness Challenges to the Computer Fraud and Abuse Act*, 94 MINN. L. REV. 1561, 1563-65 (2010) [hereinafter Kerr, *Vagueness*].

[28] 18 U.S.C. § 1030(a)(2) (2012).

[29] Defined in earlier versions as a computer used in whole or part by a financial institution or the United States government, or a crime affecting multiple computers in multiple states. *See* 18 U.S.C. § 1030(e)(2) (1987).

[30] By then the CFAA had already extended to cover computers used in all interstate and foreign commerce, and expanded crimes to cover unauthorized access to any kind of information involved in interstate and foreign communication. 18 U.S.C. §§ 1030(a)(2)(C), (e)(2)(B) (1996); *see* Kerr, *Vagueness*, *supra* note 27, at 1566–67. In 2001 the statute had broadened to cover access to foreign computers. 18 U.S.C. § 1030(e)(2)(B) (2001). In 2008 the statute's scope grew to cover access to even more forms of information. *See* Kerr, *Vagueness*, *supra* note 27, at 1569.

[31] Some of the more famous prosecutions of scraping under the CFAA include the prosecution of Aaron Swartz, United States v. Swartz, No. 11-cr-10260 (D. Mass. filed July 14, 2011), Andrew Auernheimer, United States v. Auernheimer, 748 F.3d 525 (3d Cir. 2014), the company TomorrowNow, Press Release, *TomorrowNow, Inc., Sentenced on Computer*

CFAA's corresponding civil provisions. These allow aggrieved parties to file a lawsuit to obtain damages and injunctive relief, so long as they show that they suffered a loss during a one-year period aggregating to at least $5000 in value.[32] As this loss calculation has included expenses like personnel time spent to determine the nature and extent of a scraper's activity, and possibly even money spent to hire an attorney to look into a CFAA claim,[33] it is almost always met.[34]

And so both web scraping and lawsuits about web scraping have become more common — so much so that in one current case, each side has now brought CFAA claims against the other for scraping its site.[35] But at the same time, practical advice on the legality of web scraping is hard to come by,[36] and

Intrusion and Copyright Infringement Charges, FBI (Sept. 14, 2011), https://archives.fbi.gov/archives/sanfrancisco/press-releases/2011/tomorrownow-inc.-sentenced-on-computer-intrusion-and-copyright-infringement-charges [hhtps://perma.cc/85UK-HA6N], and most recently, the founder of Oilpro.com, Press Release, *Oilpro.com Founder Sentenced to Prison for Hacking into Competitor's Computer System*, DEP'T OF JUSTICE (Oct. 6, 2017), https://www.justice.gov/usao-sdny/pr/oilprocom-founder-sentenced-prison-hacking-competitor-s-computer-system [https://perma.cc/XK7L-PQ2V]. A case was also brought against the creators of the Wiseguys ticket purchasing assistant, though the tool in this case appears to help users rapidly complete an online transaction rather than extract information. *See* United States v. Lowson, 10-cr-114, 2010 WL 9552416 (D.N.J. Oct. 12, 2010).

[32] 18 U.S.C. § 1030(g) (2012) ("A civil action for violation of this section may be brought only if the conduct involves 1 of the factors set forth in subclauses (I), (II), (III), (IV), or (V) of subsection (c)(4)(A)(i)."). This is just one possible ground for a civil action. The others, which do not appear to arise in any web scraping cases to date, address modification of medical equipment, physical injury, a threat to public health or safety, or damage affecting a federal computer used in certain key security and administration of justice functions. *Id*. § 1030(c)(4)(A)(i).

[33] Case law on this point is contradictory. The court in *Facebook, Inc. v. Power Ventures, Inc*. recently allowed recovery of attorney fees based on prior unchallenged conclusions to that effect in earlier proceedings in the case, and an independent reading of the definition of "loss" under the CFAA. 252 F. Supp. 3d 765, 777–78 (N.D. Cal. 2017), *appeal filed* No. 17-16161 (9th Cir. June 2, 2017). Other cases exclude attorney fees, or at least those incurred during the litigation itself. Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey, 497 F. Supp. 2d 627, 647–48 (E.D. Pa. 2007).

[34] *See, e.g*., Frisco Med. Ctr., L.L.P. v. Bledsoe, 147 F. Supp. 3d 646, 660 (E.D. Tex. 2015); *but see* Citizens Info. Assocs. v. Justmugshots.com, No. 1-12-CV-573-LY, 2013 WL 12076563, at *4 (W.D. Tex. Feb. 26, 2013) (finding claim of loss in that case "entirely conclusory").

[35] *See* DHI Grp., Inc. v. Kent, No. CV H-16-1670, 2017 WL 4837730, at *3-4 (S.D. Tex. Oct. 26, 2017).

[36] Some of the most helpful pieces advising scrapers include Esha Bhandari & Rachel Goodman, Data Journalism and the Computer Fraud and Abuse Act: Tips for Moving Forward in an Uncertain Landscape (Nw. Computation+Journalism Symposium, 2017) and James Snell & Nicola Menaldo, *Web Scraping in an Era of Big Data 2.0*, BLOOMBERG LAW

rarely extends beyond a rough combination of "try not to get caught" and "talk to a lawyer."[37] Most often the legal status of scraping is characterized as something just shy of unknowable, or a matter entirely left to the whims of courts, plaintiffs, or prosecutors.[38] This legal uncertainty leads to confusion and disarray on the ethical side as well, as researchers and academic publishers struggle with how to approach scraper-based research that may or not have broken a law.[39] There is also a relatively small amount of legal scholarship that addresses web scraping,[40] and the most directly on point emphasizes that the "legal doctrines involved in scraping suits are in flux."[41]

Uncertainty does indeed exist in the caselaw, and may stem in part from how courts approach the act of web scraping on a technical level. The few courts that go beyond analogies to the physical world usually describe web scraping as being akin to the actions of a human web browser, but at a far faster rate.[42] This description risks misstating the act of web scraping in a way that could affect the outcome of CFAA cases. The first goal of this piece, in Section II, is to clarify how web scrapers operate, and explain why one should not think of web scraping as being inherently more burdensome or invasive than humans browsing the web.

(June 8, 2016), https://www.bna.com/web-scraping-era-n57982073780/ [https://perma.cc/6NQA-LJAM].

[37] Some of the analysis, especially in non-legal literature, can be far worse, at best underinclusive and at worst simply wrong. *See* DAVID GOURLEY & RYAN TOTTY, HTTP: THE DEFINITIVE GUIDE 218 (2002) (noting that taking down a website can be grounds for legal claims, but without discussion of the legal concerns around accessing a website in the first place); KEVIN HEMENWAY & TARA CALISHAIN, SPIDERING HACKS 17 (2004) (suggesting that scraping won't cause liability so long as "your spiders are behaving and your intent is fair"); MITCHELL, *supra* note 11, at vii ("Some people aren't sure if it's legal (it is)[.]").

[38] *See* Jeffrey K. Hirschey, *Symbiotic Relationships: Pragmatic Acceptance of Data Scraping*, 29 BERKELEY TECH. L.J. 897, 897 (2014) (describing the "the already uncertain legal background of scraping case law"); Rami Essaid, *Is Web Scraping Illegal? It Depends on what the Meaning of the Word Is Is*, DISTIL NETWORKS, https://resources.distilnetworks.com/all-blog-posts/is-web-scraping-illegal-depends-on-what-the-meaning-of-the-word-is-is [https://perma.cc/E4NY-AZ54] (last visited July 28, 2018); Gold & Latonero, *supra* note 22, at 296.

[39] *See* Amy Bruckman, *Do Researchers Need to Abide by Terms of Service (TOS)? An Answer*, NEXT BISON (Feb. 26, 2016), https://nextbison.wordpress.com/2016/02/26/tos/ [perma.cc/76JW-EXQG].

[40] *E.g.,* Christine G. Davik, *Access Denied: Improper Use of the Computer Fraud and Abuse Act to Control Information on Publicly Accessible Internet Websites*, 63 MD. L. REV. 320 (2004); Niva Elkin-Koren, *Let the Crawlers Crawl: On Virtual Gatekeepers and the Right to Exclude Indexing*, 49 J. COPYRIGHT SOC'Y U.S.A. 165 (2001); Hirschey, *supra* note 39; Nicholas A. Wolfe, *Using the Computer Fraud and Abuse Act to Secure Public Data Exclusivity*, 13 NW. J. TECH. & INTELL. PROP. 301 (2015).

[41] Hirschey, *supra* note 39, at 926.

[42] *See infra* notes 86–92 and accompanying text.

The second goal of this piece, in Section III, is to more fully articulate how courts approach the all-important question of whether a web scraper accesses a website without authorization under the CFAA. I aim to suggest here that there is a fair amount of madness in the caselaw, but not without some method. Specifically, this piece breaks down the twenty years of web scraping litigation (and the sixty-one opinions[43] that this litigation has generated) into four rough

[43] The cases identified, in chronological order, are eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058 (N.D. Cal. 2000); Register.com, Inc. v. Verio, Inc., 126 F. Supp. 2d 238 (S.D.N.Y. 2000); EF Cultural Travel BV v. Explorica, Inc., 274 F.3d 577 (1st Cir. 2001); EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58 (1st Cir. 2003); Register.com, Inc. v. Verio, Inc., 356 F.3d 393 (2d Cir. 2004); Sw. Airlines Co. v. Farechase, Inc., 318 F. Supp. 2d 435 (N.D. Tex. 2004); Traveljungle v. Am. Airlines, Inc., 212 S.W.3d 841 (Tex. Ct. App. 2006); Facebook, Inc. v. ConnectU LLC, 489 F. Supp. 2d 1087 (N.D. Cal. 2007); Healthcare Advocates, Inc. v. Harding, Early, Follmer & Frailey, 497 F. Supp. 2d 627 (E.D. Pa. 2007); Tamburo v. Dworkin, No. 04-cv-3317, 2007 WL 3046216 (N.D. Ill. Oct. 9, 2007); Ticketmaster LLC v. RMG Techs., Inc., 507 F. Supp. 2d 1096 (C.D. Cal. 2007); A.V. v. iParadigms, LLC, 544 F. Supp. 2d 473 (E.D. Va. 2008); A.V. ex rel. Vanderhye v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009); Facebook, Inc. v. Power Ventures, Inc., No. 08-cv-5780, 2009 WL 1299698 (N.D. Cal. May 11, 2009); CollegeSource, Inc. v. AcademyOne, Inc., No. 08-cv-1987, 2009 WL 2705426 (S.D. Cal. Aug. 24, 2009); Tamburo v. Dworkin, 601 F.3d 693 (7th Cir. 2010); Snap-On Bus. Sols., Inc. v. O'Neil & Assocs., Inc., 708 F. Supp. 2d 669 (N.D. Ohio 2010); Facebook, Inc. v. Power Ventures, Inc., No. 08-cv-5780, 2010 WL 3291750 (N.D. Cal. July 20, 2010); Snapt Inc. v. Ellipse Comm'ns Inc., No. 09-cv-661, 2010 WL 11542003 (N.D. Tex. Aug. 10, 2010); Oracle USA, Inc. v. Rimini St., Inc., No. 10-cv-106, 2010 WL 3257833 (D. Nev. Aug. 13, 2010); Oracle Corp. v. SAP AG, 734 F. Supp. 2d 956 (N.D. Cal. 2010); Cvent, Inc. v. Eventbrite, Inc., 739 F. Supp. 2d 927 (E.D. Va. 2010); Snapt Inc. v. Ellipse Comm'ns Inc., No. 09-cv-661, 2010 WL 11542004 (N.D. Tex. Sept. 28, 2010); Tamburo v. Dworkin, No. 04-cv-3317, 2010 WL 5476780 (N.D. Ill. Dec. 29, 2010); Snapt Inc. v. Ellipse Comm'ns Inc., 430 Fed. App'x 346 (5th Cir. 2011); CollegeSource, Inc. v. AcademyOne, Inc., 653 F.3d 1066 (9th Cir. 2011); VRCompliance LLC v. HomeAway, Inc., No. 11-cv-1088, 2011 WL 6779320 (E.D. Va. Dec. 27, 2011); Facebook, Inc. v. Power Ventures, Inc., 844 F. Supp. 2d 1025 (N.D. Cal. 2012); EarthCam, Inc. v. Oxblue Corp., No. 11-cv-2278, 2012 WL 12836518 (N.D. Ga. Mar. 26, 2012); Dream Marriage Grp., Inc. v. Anastasia Int'l, Inc., No. 10-cv-5034, 2012 WL 3024227 (C.D. Cal. July 23, 2012); CollegeSource, Inc. v. AcademyOne, Inc., No. 10-cv-3542, 2012 WL 5269213 (E.D. Pa. Oct. 25, 2012); United States v. Auernheimer, No. 11-cr-470, 2012 WL 5389142 (D.N.J. Oct. 26, 2012); Citizens Info. Assocs. v. Justmugshots.com, No. 12-cv-573, 2012 WL 12874898 (W.D. Tex. Dec. 18, 2012); Citizens Info. Assocs. v. Justmugshots.com, No. 12-cv-573, 2013 WL 12076563 (W.D. Tex. Feb. 26, 2013); Craigslist Inc. v. 3Taps. Inc., 942 F. Supp. 2d 962 (N.D. Cal. 2013); VRCompliance LLC v. HomeAway, Inc., 715 F.3d 570 (4th Cir. 2013); EarthCam, Inc. v. Oxblue Corp., 11-cv-2278, 2013 WL 11904713 (N.D. Ga. July 19, 2013); Craigslist Inc. v. 3Taps Inc., 964 F. Supp. 2d 1178 (N.D. Cal. 2013); Facebook, Inc. v. Power Ventures, Inc., 08-cv-5780, 2013 WL 5372341 (N.D. Cal. Sept. 25, 2013); Tamburo v. Dworkin, 974 F. Supp. 2d 1199 (N.D. Ill. 2013); Fidlar Techs. v. LPS Real Estate Data Sols., Inc., No. 13-cv-4021, 2013 WL 5973938 (C.D. Ill. Nov. 8, 2013); United States v. Auernheimer, 748 F.3d 525 (3d Cir. 2014); EarthCam, Inc. v. Oxblue Corp., 49 F. Supp. 3d 1210 (N.D. Ga. 2014); CollegeSource, Inc. v. Acade-

phases of thinking around the critical question of when a scraper access a computer "without authorization" or if it "exceeds authorized access."

The first phase runs through the first decade of scraping litigation, and is marked with cases that adopt an expansive interpretation of the CFAA, with the potential to extend to all scrapers, so long as a website can point to some mechanism to signal that the access was unauthorized, be that contractual, technical, or otherwise.[44] In the second phase, starting in the late 2000s and following an influential wave of cases which began to adopt a "narrow" view of the CFAA,[45] courts began to deny claims in scraping cases where websites merely placed restrictions on the *use* of the data hosted on their site as opposed to restrictions on *access* to a website, and looked more towards code-based controls to interpret the scope of a scraper's authorization.[46] This approach

myOne, Inc., 597 Fed. App'x 116 (3d Cir. 2015); Fidlar Techs. v. LPS Real Estate Data Sols., Inc., 82 F. Supp. 3d 844 (C.D. Ill. 2015); QVC, Inc. v. Resultly, LLC, 99 F. Supp. 3d 525 (E.D. Pa. 2015); CollegeSource, Inc. v. AcademyOne, Inc., 08-cv-1987, 2015 WL 5638104 (S.D. Cal. Sept. 24, 2015); Fidlar Techs. v. LPS Real Estate Data Sols., Inc., 810 F.3d 1075 (7th Cir. 2016); QVC, Inc. v. Resultly, LLC, 159 F. Supp. 3d 576 (E.D. Pa. 2016); CouponCabin LLC v. Savings.com, Inc., No. 14-cv-39, 2016 WL 3181826 (N.D. Ind. June 8, 2016); Oracle USA, Inc. v. Rimini St., Inc., 191 F. Supp. 3d 1134 (D. Nev. 2016); Facebook, Inc. v. Power Ventures, Inc., 828 F.3d 1068, *opinion superseded on denial of reh'g en banc* 844 F.3d 1058 (9th Cir. 2016); CouponCabin LLC v. Savings.com, Inc., No. 14-cv-39, 2017 WL 83337 (N.D. Ind. Jan. 10, 2017); DHI Grp., Inc. v. Kent, No. 16-cv-1670, 2017 WL 1088352 (S.D. Tex. Mar. 3, 2017); Heritage Capital Corp. v. Christie's Inc., No. 16-cv-3404, 2017 WL 1550514 (N.D. Tex. May 1, 2017); Facebook, Inc. v. Power Ventures, Inc., 252 F. Supp. 3d 765 (N.D. Cal. 2017); EarthCam, Inc. v. Oxblue Corp., 703 Fed. App'x 803 (11th Cir. 2017); hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099 (N.D. Cal. 2017); DHI Grp., Inc. v. Kent, No. 16-cv-1670, 2017 WL 4837730 (S.D. Tex. Oct. 26, 2017); Oracle USA, Inc. v. Rimini St., Inc., 879 F.3d 948 (9th Cir. 2018); Sandvig v. Sessions, No. 16-cv-1638 (JDB), 2018 WL 1568881 (D.D.C. March 30, 2018). As noted below, a little less than two-thirds of these go beyond procedural issues or passing mentions of CFAA claims to actually examine the substance of how the law applies. This list also does not consider cases concerning the related-but-distinct technique of using automated scripts to complete online transactions. *See, e.g.*, Ticketmaster L.L.C. v. Prestige Entm't, Inc., No. 17-cv-7232, 2018 WL 654410 (C.D. Cal. Jan. 31, 2018) (defendant wrote script to rapidly acquire large numbers of event tickets); Craigslist, Inc. v. Naturemarket, Inc., 694 F. Supp. 2d 1039 (N.D. Cal. 2010) (defendant wrote scripts to batch-post material onto online classifieds service). It also does not look at cases where automatic requests for websites were used to deliberately overwhelm a website server in a "denial of service attack." *See, e.g.*, United States v. Yücel, 97 F. Supp. 3d 413 (S.D.N.Y. 2015).

[44] *See infra* Section III.A.

[45] Jonathan Mayer, *The "Narrow" Interpretation of the Computer Fraud and Abuse Act: A User Guide for Applying* United States v. Nosal, 84 GEO. WASH. L. REV. 1644, 1657 (2016) [hereinafter Mayer, *The "Narrow" Interpretation*]. Notable cases include *LVRC Holdings LLC v. Brekka*, 581 F.3d 1127 (9th Cir. 2009) and *United States v. Nosal (Nosal I)*, 676 F.3d 854 (9th Cir. 2012) (en banc). For further analysis of this trend, *see, e.g.*, Wolfe, *supra* note 40.

[46] *See infra* Section III.B.

tended to benefit web scrapers, because (as noted in Section II) in most cases a scraper confronts no further code-based restriction than a human would at a web browser, and barriers like a website's terms of use tend to limit only the use of information, rather than access to information, making them unsuitable grounds for a CFAA claim under this newly-narrowed view.[47]

But in a shift that has gone less observed, a third phase of analysis has grown over the last half-decade, which brings interpretation of the CFAA back into the older, broader view. This change was brought about in part by a reexamination of the Ninth Circuit's landmark 2009 case *LVRC Holdings LLC* v. *Brekka*.[48] The case is typically thought of as a hallmark case for the "narrow" view of the CFAA — and in in the scraping world it was, at first.[49] But starting in 2013 courts began to look to other language in *Brekka* to develop a new "revocation" theory under the CFAA, where a website could establish liability if it could show that at some point the site "revoked" access to the scraper, and the scraper continued to access the site.[50] And instead of carefully examining the language of a restriction or looking solely to technical controls, courts allowed claims based on mechanisms that arguably "revoked" access, and thus reintroduced CFAA claims hinging on less-concrete factors, such as the contents of a website terms of use, a direct demand to stop access a public website, a scraper's implied knowledge of third-party contracts, and even the use of a technical block without any notice or other communication to the scraper.[51] In light of this shift, prior resistance to applying the CFAA as a means of enforcing restrictions on "use" of content was irrelevant; so long as a website could show that it acted upon its objections by completely revoking the scraping party's access, the site could invoke the CFAA.[52]

---

[47] This is especially true given some courts' careful analysis of whether a purported "access restriction" is just a "use restriction" in disguise. *See*, *Wentworth-Douglass Hosp. v. Young & Novis Prof. Ass'n,* No. 10-cv-120-SM, 2012 WL 2522963, at *4 (D.N.H. June 29, 2012) ("[D]enominating limitations as 'access restrictions' does not convert what is otherwise a use policy into an access restriction").

[48] 581 F.3d 1127. *See infra* notes 210–218 and accompanying text.

[49] *See* Cvent, Inc. v. Eventbrite, Inc., 739 F. Supp. 2d 927, 933 (E.D. Va. 2010) (citing *Brekka* as supporting the district's earlier decision to limit claims based on use restrictions instead of access restrictions); Facebook, Inc. v. Power Ventures, Inc., No. C 08-05780 JW, 2010 WL 3291750, at *9 (N.D. Cal. July 20, 2010) (citing *Brekka* specifically to find that "more recent CFAA cases militate for an interpretation [of the California CFAA equivalent] that does not premise permission to access or use a computer or computer network on a violation of terms of use").

[50] This started with *Craigslist Inc. v. 3Taps Inc.*, 964 F. Supp. 2d 1178 (N.D. Cal. 2013), and reached highest prominence with *Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058 (9th Cir. 2016).

[51] *See infra* notes 234–235 and accompanying text.

[52] *Craigslist*, 964 F. Supp. 2d at 1185 (finding the fact that a website brought an action against a scraper because of how the scraper used the information obtained to be "true, but beside the point").

Most recently, spurred in part by the same policy concerns that led courts to initially constrain the CFAA in the first place,[53] courts have begun rethinking this result. Three opinions issued in the past few months[54] have begun to reject this broader reading, either by finding a different rule for public websites or by more strictly defining what constitutes "revocation."[55] These opinions either do not address the "revocation" theory or purport to fit their analysis into the "revocation" line of cases identified above, but in a way that would seem to make it far more difficult to stop a scraper from accessing a website available to the general public, even if told to stop by the website in question. To the most recent court to address the question, scraping "is merely a technological advance that makes information collection easier," and if human user can collect information on the Internet, a scraper can, too.[56]

In sum, there is a pattern as to how courts have approached application of the CFAA to web scraping. There has been a subtle evolution in thinking that has worked its way into the two decades of CFAA case law, albeit one that that has at various times given differing levels of clarity to scrapers who seek to understand whether their activity violates this law. The conclusion of this piece identifies broader questions about the CFAA and web scraping which courts must address in order to bring more harmony and comprehension to this area of law. Those questions include how to deal with conflicting instructions on authorization coming from different channels on the same website; how the analysis should interact with the existing technical protocols that regulate web scraping, including the Robots Exclusion Standard; and, beyond the interests of the website host, what other factors should govern application of the CFAA to unwanted web scraping of public websites.

II. REFINING THE TECHNICAL EXPLANATION OF INTERNET SCRAPING

At the outset, it is worth taking some time to more precisely define what is meant by "web scraping." Courts have struggled to settle on a common terminology for web scraping, let alone what types of activity should meet the definition.[57] They have used terms ranging from "scraping programs,"[58] "screen

[53] *See*, hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099, 1110–11 (N.D. Cal. 2017).

[54] Oracle USA, Inc. v. Rimini Street, Inc., 879 F.3d 948 (9th Cir. 2018); Sandvig v. Sessions, No. 16-cv-1638 (JDB), 2018 WL 1568881 (D.D.C. Mar. 30, 2018); *hiQ Labs*, 273 F. Supp. 3d. 1099.

[55] *See infra* Section III.D.

[56] *Sandvig*, 2018 WL 156881 at *7.

[57] *See* eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1060 n.2 (N.D. Cal. 2000) ("Programs that recursively query other computers over the Internet in order to obtain a significant amount of information are referred to in the pleadings by various names, including software robots, robots, spiders and web crawlers.").

[58] CouponCabin LLC v. Savings.com, Inc., No. 2:14-cv-39-TLS, 2016 WL 3181826, at *1 (N.D. Ind. June 8, 2016).

scraping,"[59] "a robot web crawling program,"[60] or use of a "robot,"[61] "automatic web browser,"[62] "webcrawlers,"[63] "spider,"[64] or, confusingly, a "search engine."[65] Some courts attempt to differentiate between these terms based on how many websites are targeted,[66] how much is copied,[67] or by different steps in the process of data extraction.[68] Those who scrape have been viewed as anything from a vital public benefactor[69] to, in the colorful words of one objecting party, "a low lying snake belly scum sucking rat" who should be "quartered and hung."[70]

Analogies and metaphors permeate the opinions as well, though they seem to generate more confusion than they remedy.[71] Websites are often likened by litigants and courts to stores,[72] though sometimes instead a bank,[73] a fruit

[59] Dream Marriage Grp., Inc. v. Anastasia Int'l, Inc., No. CV 10-5034 RSWL (FFMX), 2012 WL 3024227, at *1 (C.D. Cal. July 23, 2012).

[60] QVC, Inc. v. Resultly, LLC, 159 F. Supp. 3d 576, 589 (E.D. Pa. 2016).

[61] EF Cultural Travel BV v. Explorica, Inc., 274 F.3d 577, 579 (1st Cir. 2001).

[62] Internet Archive v. Shell, 505 F. Supp. 2d 755, 760 (D. Colo. 2007).

[63] Ticketmaster Corp. v. Tickets.com, Inc., No. 99CV7654, 2000 WL 1887522, at *2 (C.D. Cal. 2000).

[64] Ticketmaster Corp. v. Tickets.com, Inc., No. CV997654HLHBKX, 2003 WL 21406289, at *2 (C.D. Cal. Mar. 7, 2003).

[65] Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1155 (9th Cir. 2007) ("Google operates a search engine, a software program that automatically accesses thousands of websites[.]").

[66] EF Cultural Travel BV v. Explorica, Inc., 274 F.3d 577, 579 (1st Cir. 2001) (distinguishing between "robots," which "gather information for countless purposes" across many websites, and a "scraper" who is "focused solely on [plaintiff's] website").

[67] *See* Hirschey, *supra* note 40, at 898 (noting that search engines are usually referred to as "crawlers," while more invasive retrieval tools are called "scrapers.").

[68] Compulife Software, Inc. v. Newman, No. 9:16-CV-81942-ROSENBERG/BRANNON, 2017 WL 2537357, at *4 (S.D. Fla. June 12, 2017) (referring to scraping as inserting data into scraper's database); CollegeSource, Inc. v. AcademyOne, Inc., No. 10-3542, 2012 WL 5269213, at *4 (E.D. Pa. Oct. 25, 2012) (noting that downloaded .pdf files have to be "converted and processed before being copied as text, or 'scraped,' into [defendant's] database").

[69] A.V. *ex rel*. Vanderhye v. iParadigms, LLC, 562 F.3d 630, 638 (4th Cir. 2009).

[70] Tamburo v. Dworkin, 974 F. Supp. 2d 1199, 1210 (N.D. Ill. 2013).

[71] Some courts make a conscious point to distance themselves from such analogies. United States v. Auernheimer, 748 F.3d 525, 541 (3d Cir. 2014) ("[W]e must remain mindful that cybercrimes do not happen in some metaphysical location . . . . People and computers still exist in identifiable places in the physical world."); Fidlar Techs. v. LPS Real Estate Data Sols., Inc., 82 F. Supp. 3d 844, 855–56 (C.D. Ill. 2015) (criticizing one litigant's theory of a computer system as "almost metaphysical in its abstraction").

[72] *See, e.g.*, Craigslist Inc. v. 3Taps Inc., 964 F. Supp. 2d 1178, 1187 (N.D. Cal. 2013); eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1065 (N.D. Cal. 2000). This analogy finds its way into scholarship as well. *See* Orin S. Kerr, *Cybercrime's Scope: Interpreting "Access" and "Authorization" in Computer Misuse Statutes*, 78 N.Y.U. L. REV. 1596,

stand,[74] a food truck,[75] or a bulletin board.[76] Scrapers have in turn been likened to an invading army of robots,[77] a vandal taking hammer to a piece of machinery,[78] a person walking into a bank with both a safety deposit key and a shotgun[79] — or, more innocently, a roving machine that constantly takes photographs,[80] an interviewer using an audio recording instead of taking notes,[81] or a person who records signs posted within a store.[82] It is hard to see what guiding principles one can draw from such an array of conflicting imagery.

The Southern District of New York appears to have been the first court to define a web scraper in 1996, as "software capable of automatically contacting various Web sites and extracting relevant information."[83] This definition has an elegant structure, but upon closer examination becomes over-inclusive. After all, web browsers like Firefox or Chrome are also capable of automatically contacting websites to extract information. The process of loading a modern website necessarily requires the browser to contact the numerous additional other servers that host the underlying images, banner ads, social media buttons, tracking pixels, and other objects.[84] More recently browsers have also begun "link prefetching," or loading pages that are linked off of the page that the user most recently loaded.[85] Both link-prefetching and the modern web browser functionality explained above could theoretically fall under the Southern District's definition, but it is highly unlikely that most lawyers or coders would consider them to be "web scraping."

1620 (2003) ("[W]e could say that visiting a publicly accessible website is something like visiting an open store in the physical world.") [hereinafter Kerr, *Cybercrime's Scope*].

[73] Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1068 (9th Cir. 2016).

[74] Register.com, Inc. v. Verio, Inc., 356 F.3d 393, 401 (2d Cir. 2004).

[75] Sandvig v. Sessions, No. 16-cv-1638 (JDB), 2018 WL 1568881 at *5 (D.D.C. Mar. 30, 2018).

[76] Facebook, Inc. v. ConnectU LLC, 489 F. Supp. 2d 1087, 1093 (N.D. Cal. 2007).

[77] *eBay*, 100 F. Supp. 2d at 1065. The court, to its credit, said that "[t]his analogy, while graphic, appears inappropriate." *Id*.

[78] Ticketmaster Corp. v. Tickets.com, Inc., No. 99CV7654, 2000 WL 1887522, at *4 (C.D. Cal. Aug. 10, 2000) (order denying preliminary injunction).

[79] Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1068 (9th Cir. 2016).

[80] Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey, 497 F. Supp. 2d 627, 631 (E.D. Pa. 2007).

[81] *Sandvig*, 2018 WL 1568881 at *7.

[82] hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099, 1112–13 (N.D. Cal. 2017).

[83] Shea *ex rel*. The Am. Reporter v. Reno, 930 F. Supp. 916, 929 (S.D.N.Y. 1996) (discussing early search engines).

[84] MITCHELL, *supra* note 11, at 5; *see infra* notes 103–104 and accompanying text.

[85] Perhaps significantly, this is usually done with explicit instructions written into in the HTML by the website host. *See* Addy Osmani, *Preload, Prefetch, and Priorities in Chrome*, MEDIUM (Mar. 27, 2017), https://medium.com/reloading/preload-prefetch-and-priorities-in-chrome-776165961bbf [https://perma.cc/Q8SW-U9M4].

A more detailed definition comes from the First Circuit in the 2003 scraping case *EF Cultural Travel BV v. Zefer Corp.*:

> A scraper, also called a "robot" or "bot," is nothing more than a computer program that accesses information contained in a succession of webpages stored on the accessed computer. Strictly speaking, the accessed information is not the graphical interface seen by the user but rather the HTML source code — available to anyone who views the site — that generates the graphical interface. This information is then downloaded to the user's computer.[86]

This definition draws closer to the mark, but overlooks some details. First, many applications of scraping don't require the retrieval of a succession of pages on the same computer. They could instead look to follow links around to other websites hosted on other computers.[87] Second, material on most modern websites is rarely statically "stored" on pages, patiently waiting to be "extracted" by a scraper. Most websites instead are dynamically generated as they are requested, often drawing upon information provided by the user seeking access, including the user's account information, time of day, geographic location, and whether the user is accessing the page from a mobile device.[88] Indeed, one application of web scraping is to provide insight into how these inputs can change the outputs. For example, one recent research project looked at how prices for online products varied based on the user's location by varying the reported zip code of the scraper to see whether the retailer provided differing prices for the same products (as indeed they did).[89]

On a broader level, courts can also run astray if they start their analysis at what a human sees at the web browser level and work from there to get to the data that scrapers extract,[90] or imagine the scraper as an automaton replicating the steps of a human at a faster rate.[91] This approach can make it seem as

[86] 318 F.3d 58, 60 (1st Cir. 2003).

[87] HEMENWAY & CALISHAIN, *supra* note 37, at 150–51.

[88] *See* GOURLEY & TOTTY, *supra* note 37 at 4–5.

[89] Katja Seim & Michael Sinkinson, *Mixed Pricing in Online Marketplaces*, 14 QUANT. MARKETING & ECON. 129, 131 (2016).

[90] *See* Nautical Sols. Mktg., Inc. v. Boats.com, No. 8:02–CV–760–T–23TGW, 2004 WL 783121, at *1 (M.D. Fla. Apr. 1, 2004) (a scraper "visits targeted public websites, extracts facts from the websites and indexes the extracted facts . . . ."); Ticketmaster Corp. v. Tickets.com, Inc., No. CV997654HLHVBKX, 2003 WL 21406289, at *2 (C.D. Cal. Mar. 7, 2003) ("The 'spider' 'crawled' through the internal web pages . . . and electronically extracted the electronic information from which the web page is shown on the user's computer.").

[91] *See, e.g.*, *In re* Complaint of Judicial Misconduct, 575 F.3d 279, 288 (3d Cir. 2009) (defining crawlers as "sophisticated automated web-scanning software . . . that aggressively catalogues and indexes website content"); Snap-On Bus. Sols., Inc. v. O'Neil & Assocs., Inc., 708 F. Supp. 2d 669, 674 (N.D. Ohio 2010) (quoting from defendant's testimony, describing scraping as "simulat[ing] what a user could do interactively with the website by

though a website scraper is an elaborate layer on top of a web browser, perhaps adding more of a burden on the website or going deeper than a normal web browser could. Most scrapers operate instead on a simpler level, and retrieve the objects and files used to build a visible webpage *before* they are rendered and displayed to the user.[92]

The confusion here could stem from the layer of Internet architecture on which courts tend to focus their analyses. As has long been observed in the field of Internet design, the structure of the Internet resembles a layered hourglass, with different layers representing different aspects of network and communications architecture, a principal genius of the Internet's architecture being the fact that, in the words of Prof. Jonathan Zittrain, "[t]inkerers can work on one layer without having to understand much about the others, and there need not be any coordination or relationship between those working at one layer and those at another."[93]

Near the top of this hourglass are the layers that an average computer user thinks about, things like web browsers and the Hypertext Markup Language (HTML) code that creates a webpage.[94] One layer down from this is the Hypertext Transfer Protocol (HTTP), which is the layer at which most web scrapers operate.[95] HTTP is the protocol by which all traffic on the World Wide Web is formatted for communication,[96] and addresses how all media, pages, scripts, and other files (referred to generally as "web resources") are created, stored, and retrieved on web servers.[97] The protocol defines the roles of a "server," or the computer generating and hosting web resources, and a "client," the requester of web resources.[98] The protocol also defines the commands a client can use to request information from a server — including GET, to re-

pointing and clicking, only it's automated, and, therefore, able to point, click and do other things that the user would do in an automated manner, making it able to run unattended in a much more efficient way"); Playboy Enters., Inc. v. Terri Welles, Inc., 78 F. Supp. 2d 1066, 1092 (S.D. Cal. 1999), *rev'd*, 279 F.3d 796 (9th Cir. 2002) ("Web crawlers 'read' individual web pages by reading much of the text in the HTML source code and store in cyberspace memory the text they find on each page.").

[92] GOURLEY & TOTTY, *supra* note 37, at 8–9. There are some web scrapers that are designed instead to operate on top of a standard web browser. *See, e.g.*, WEBSCRAPER, http://webscraper.io/ (last visited Mar. 6, 2018) [https://perma.cc/2W4V-5B8W] (a browser extension for the Chrome browser to enable web scraping).

[93] JONATHAN ZITTRAIN, THE FUTURE OF THE INTERNET (AND HOW TO STOP IT) 68 (2008).

[94] I am being a bit loose with defining the layers because, as Zittrain notes, "[t]he exact number of layers varies depending on who is drawing the hourglass and why . . . ." *Id*. at 67.

[95] MITCHELL, *supra* note 11, at 178–80 (noting various ways scrapers can tinker with settings at the HTTP layer to achieve different results).

[96] GOURLEY & TOTTY, *supra* note 37, at 3.

[97] *Id*. at 4.

[98] *Id*. To be slightly more specific, when generating content, a server may also act as a "gateway," converting HTTP traffic into another protocol for another computer to process and respond to. *See id*. at 19.

trieve a particular resource; POST, to send client data to a server application; and HEAD, to send just the metadata (the "HTTP headers") around a particular resource from the server to the client.[99]

Programs used by clients to retrieve web resources from servers are known as "user agents."[100] A web browser is one form of user agent. A scraper is another. Either way, a standard communication between a server and a user agent will start with the agent making a request to the server for particular information, including the method of communication (usually GET), the address of the requested information, and various "headers" that may contain additional information relevant to the request, such as the requester's operating system, Internet Protocol (IP) address, or the address of the website the agent came from.[101] The server will take that information and use it to formulate an appropriate response, and then send the requested data.[102] So when loading a webpage in a web browser the user agent (in that case, a web browser) sends an HTTP request to the server to obtain the HTML file that sets forth the content and layout of the webpage. It then issues multiple additional HTTP transactions with the same server (and likely other servers) to build the various other elements that constitute the web page's contents: a banner ad here, an embedded social media post there, and so forth.[103] Users rarely notice this happening, but some web browsers allow you to log these transactions to see this cascade play out.[104]

At this layer a scraper works in the same way a web browser does.[105] It sends out HTTP transactions for the web resources that it seeks along the same protocols, and the server sends the same files in return.[106] The scraper's level

[99] *Id*. at 8–9. Courts occasionally find their way to discussing the various HTTP headers in the context of different cases. *See* Fidlar Techs. v. LPS Real Estate Data Sols., Inc., 810 F.3d 1075, 1077–78 (7th Cir. 2016) (in a scraping case, examining systematic "SOAP" or "Simple Object Access Protocol" requests made using the POST method); *see also*, *In re* Pharmatrak, Inc., 329 F.3d 9, 16 (1st Cir. 2003) (privacy concerns related to sending information using the GET method, versus the POST method); Level 3 Comm'ns, LLC v. Limelight Networks, Inc., 589 F. Supp. 2d 664, 674–75 (E.D. Va. 2008) (a patent *Markman* opinion concerning competing techniques for extracting information from HTTP headers).

[100] GOURLEY & TOTTY, *supra* note 37, at 19–20.

[101] *Id*. at 258–59.

[102] *Id*. at 69.

[103] *Id*. at 9.

[104] Josh Gough, *How to Spy on Your Browser's HTTP Requests and Responses!*, VERSIONONE BLOG (Feb. 7, 2013), https://blog.versionone.com/spy-on-browser-http-requests/ [https://perma.cc/Q239-4JQX].

[105] Though courts do not always appreciate that this is the case. *See, e.g.*, Compulife Software, Inc. v. Newman, No. 9:16-CV-81942-ROSENBERG/BRANNON, 2017 WL 2537357, at *3 (S.D. Fla. June 12, 2017) (referring to "get commands" as "an alternative way to communicate with the host-based software without going through a website," when, in fact, a web browser would also send GET requests to load a webpage).

[106] GOURLEY & TOTTY, *supra* note 37, at 19–20.

of access is just as deep as a web browser's, and the method by which it makes its queries is identical.[107] The principal difference between a scraper and a normal web browser is that the material presented is not rendered and presented to the user; it is instead used for some other purpose. This can mean that as to any given web page the load placed on the host's server may in fact be *lighter*, because the scraper may only need one web resource, rather than the dozens a web-browser might need, in order to extract the relevant information.[108] And while it is not required, web scrapers can include a "user agent header" in their requests, identifying the name of their scraper.[109] HTTP even allows the scraper operators to provide an email address in case a server's administrator wishes to contact them.[110]

Perhaps the largest difference between browsing and scraping is that, where browsing allows a user to "collect" the assorted contents of a particular webpage, most scrapers will collect information from a series of different webpages.[111] Indeed, the challenge in developing an effective scraper is to understand where and how the data in question is built and stored, so one can write a scraping program that will retrieve the greatest amount of desired information and the least amount of noise.[112] Also, because scrapers often request pages serially, a misconfigured scraper can get caught up into accidental "loops" and "dups" based on how the server responds to the scraper's request.[113] In those situations scrapers risk overwhelming a website and crashing it, and programmers have developed an array of techniques to help prevent this.[114] After all, a website scraper generally does not want the site to crash; it wants to access the site's contents.[115]

Properly contextualized, therefore, the access a server provides to a web scraper is highly similar to that provided to a standard web browser. The scraper requests and receives files using the same protocols as a web browser,

[107] Other legal scholars have observed this fact as well. Davik, *supra* note 40, at 332; Maureen O'Rourke, *Property Rights and Competition on the Internet: In Search of an Appropriate Analogy*, 16 BERKELEY TECH. L.J. 561, 570 (2001).

[108] MITCHELL, *supra* note 11, at viii–ix.

[109] *See* United States v. Auernheimer, 748 F.3d 525, 530 (3d Cir. 2014) (defendant altered the user agent on the web browser in order to match what the server expected when delivering content); QVC, Inc. v. Resultly, LLC, 99 F. Supp. 3d 525, 530 (E.D. Pa. 2015) (noting that defendant scraper identified itself in its user agent header, even though "[t]here is no requirement that [defendant] identify itself in this way"); GOURLEY & TOTTY, *supra* note 37, at 225.

[110] *Id*.

[111] *Id*. at 215.

[112] *Id*. at 223–24 (outlining strategies to that end).

[113] *Id*. at 217–18.

[114] *See, e.g.*, HEMENWAY & CALISHAIN, *supra* note 37, at 42–45.

[115] *See, e.g.*, QVC, Inc. v. Resultly, LLC, 159 F. Supp. 3d 576, 594 (E.D. Pa. 2016) (noting that it was implausible to claim that scraper intended to cause damage under the Computer Fraud and Abuse Act, as defendant relied on website to operate its business).

and at least in some cases, places less of a load on a website by only retrieving the objects necessary to extract certain information, rather than all of the material to visually render a website for a human reader. The difference between a scraper and a web browser comes less from differences in how scrapers *access* servers, and more from what the scraper does with the information after it is loaded. To return to the CFAA, to the extent the statute targets those who access websites without permission,[116] it would seem as though web scraping should rarely pose an issue under the statute — at most, particular *uses* of scraped material could be examined by other doctrines that police the use of information, such as copyright law.[117] That has not been the experience of web scraping under the CFAA, however, as detailed below.

III. PHASES OF THINKING ON WEB SCRAPING AND THE CFAA

There have been about sixty-one opinions that have considered the application of the CFAA (or state equivalents thereof) to web scraping.[118] About thirty-nine of these opinions go beyond procedural questions and other ancillary issues to directly analyze the substantive claims.[119] The opinions begin in 2000, a little less than a decade after the establishment of HTTP and the World Wide Web in 1991,[120] and grow in frequency nearly every year since, from one

[116] This would be akin to the "trespass" formulation that Prof. Orin Kerr and others have put forth as a guiding framework. *See* Orin Kerr, *Norms of Computer Trespass*, 116 COLUM. L. REV. 1143 (2016) [hereinafter Kerr, *Norms*].

[117] As indeed they have. *See, e.g.*, Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537 (S.D.N.Y. 2013) (finding copyright infringement based on a scrape); Field v. Google Inc., 412 F. Supp. 2d 1106 (D. Nev. 2006) (finding fair use based on a scrape).

[118] Claims related to web scraping have been raised under several states' computer access laws. *See* Oracle USA, Inc. v. Rimini St., Inc., 879 F.3d 948 (9th Cir. 2018) (California and Nevada); Fidlar v. LPS Real Estate Data Sols., Inc., 810 F.3d 1075 (7th Cir. 2016) (Illinois); DHI Grp., Inc. v. Kent, No. H-16-1670, 2017 WL 4837730 (S.D. Tex. Oct. 26, 2017) (Texas); United States v. Auernheimer, No. 11-cr-470 (SDW), 2012 WL 5389142 (D.N.J. Oct. 26, 2012), *rev'd on other grounds*, 748 F.3d 525 (3d Cir. 2014) (New Jersey); Earthcam, Inc. v. Oxblue Corp., No. 1:11-cv-02278-WSD, 2012 WL 12836518 (N.D. Ga. Mar. 26, 2012) (Georgia); Cvent, Inc. v. Eventbrite, Inc., 739 F. Supp. 2d 927 (E.D. Va. 2010) (Virginia). Except as otherwise noted, most courts analyze questions of website authorization under a similar framework under both federal and state computer access laws, and cases that raised state equivalents are analyzed as part of the set of opinions discussed here.

[119] *See infra* notes 159, 191, 231.

[120] *See*, *The Original HTTP as Defined in 1991*, WORLD WIDE WEB CONSORTIUM (1991), https://www.w3.org/Protocols/HTTP/AsImplemented.html. There are some older precedents to this analysis, including cases that examined scraping-like activity in the context of companies sending spam emails on the America Online and CompuServe platforms. *See, e.g.*, Am. Online, Inc. v. LCGM, Inc., 46 F. Supp. 2d 444 (E.D. Va. 1998); CompuServe Inc. v. Cyber Promotions, 962 F. Supp. 1015 (S.D. Ohio 1997). The most famous form of web scraper, the search engine, was also mentioned in earlier cases that considered the constitutional challenges to the Communications Decency Act, Am. Civil Liberties Union v. Reno,

to two opinions per year in the early 2000s[121] to closer to six to eight per year in the 2010s.[122] This roughly tracks the expansion of the CFAA in the civil context more broadly.[123] There have been a little over a dozen appellate opinions in cases involving web scraping,[124] but only one has generated something resembling a dissenting opinion.[125]

Before turning to the particular analyses, there are observations to make about the set as a whole. First, it is important to note that of the sixty-one opinions identified, about a third stem from just four underlying disputes: a decade-long litigation between Facebook and would-be social network aggregator Power.com;[126] parallel litigation in California and Pennsylvania between two rival services that assist college students who transfer schools;[127] a series of claims brought by a scraper of dog pedigree databases against data hosts who

929 F. Supp. 824 (E.D. Pa. 1996); Shea *ex rel*. The Am. Reporter v. Reno, 930 F. Supp. 916 (S.D.N.Y. 1996), and cases addressing trademark law and search indexes, Nettis Envtl. Ltd. v. IWI, Inc., 46 F. Supp. 2d 722, 727–28 (N.D. Ohio 1999); Insituform Techs., Inc. v. Nat'l Envirotech Grp., L.L.C, No. 97-2064, 1997 WL 34658315, at *2 (E.D. La. Aug. 26, 1997); Toys "R" Us, Inc. v. Akkaoui, No. C 96-3381 CW, 1996 WL 772709, at *5 (N.D. Cal. Oct. 29, 1996).

[121] *See supra* note 43.

[122] *See id*.

[123] Jonathan Mayer, *Cybercrime Litigation*, 164 U. PA. L. REV. 1453, 1476 (2016).

[124] Oracle USA, Inc. v. Rimini St., Inc., 879 F.3d 948 (9th Cir. 2018); EarthCam, Inc. v. OxBlue Corp., 703 Fed. App'x 803 (11th Cir. 2017); Facebook, Inc. v. Power Ventures, Inc., 828 F.3d 1068, *superseded by*, 844 F.3d 1058 (9th Cir. 2016); Fidlar Techs. v. LPS Real Estate Data Sols., Inc., 810 F.3d 1075 (7th Cir. 2016); CollegeSource, Inc. v. AcademyOne, Inc., 597 Fed. App'x 116 (3d Cir. 2015); United States v. Auernheimer, 748 F.3d 525 (3d Cir. 2014); VRCompliance LLC v. HomeAway, Inc., 715 F.3d 570 (4th Cir. 2013); CollegeSource, Inc. v. AcademyOne, Inc., 653 F.3d 1066 (9th Cir. 2011); Snapt Inc. v. Ellipse Commc'ns, Inc., 430 Fed. App'x 346 (5th Cir. 2011); Tamburo v. Dworkin, 601 F.3d 693 (7th Cir. 2010); A.V. *ex rel*. Vanderhye v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009); Register.com, Inc. v. Verio, Inc., 356 F.3d 393 (2d Cir. 2004); EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58 (1st Cir. 2003); EF Cultural Travel BV v. Explorica, Inc., 274 F.3d 577 (1st Cir. 2001).

[125] Judge Fred I. Parker of the Second Circuit was initially assigned to write the majority opinion in *Register.com, Inc*., 356 F.3d at 395 n.1. In the process of doing so, he changed his mind on the result, but was unable to convince Judges Leval and Keenan to join a brief that would have reversed the preliminary injunction against the scraper in question. *Id*. Judge Parker passed away before drafting a formal dissent, and the court appended a draft of his would-be majority opinion reversing the injunction to their decision upholding the injunction. *See id*. at 406. In the end, the court did not use the CFAA as grounds to issue the injunction, *id*., and Judge Parker would have vacated the injunction on both the CFAA and other grounds, *id*. at 440.

[126] Facebook, Inc. v. Power Ventures, Inc., 252 F. Supp. 3d 765 (N.D. Cal. 2017); *see supra* note 43.

[127] *CollegeSource*, 597 F. App'x 116; *see supra* note 43.

called him a criminal and thief;[128] and litigation between two rival security camera companies with entangled trade secrets issues.[129] This high concentration of opinions in a few specific cases could mean that the facts of those cases, and the courts deciding them, stand to have an outsized influence in our understanding of the doctrine to date.

Second, a tremendous number of these opinions concern claims brought by direct commercial competitors[130] or companies in closely adjacent markets to each other.[131] A far smaller number involve commercial scrapers with non-commercial hosts.[132] Only three opinions involve a commercial data host and a public-interest-oriented scraper: a declaratory action brought by an association of resort towns who used a scraping service to determine whether home rentals facilitated on an online platform were evading tax obligations;[133] a hacker who discovered a security oversight in AT&T's website for iPad users, who then gathered a list of email addresses leaked through this oversight and gave the list to the online publication Gawker;[134] and a constitutional challenge to the

[128] These cases also contribute very little in the way of analysis, as the CFAA claim was largely ancillary to other issues. *Tamburo*, 974 F. Supp. 2d 1199; *see supra* note 43.

[129] *EarthCam*, 703 F. App'x 803; *see supra* note 43.

[130] *See, e.g.*, *EarthCam*, 703 Fed. App'x 803; EF Cultural Travel BV v. Explorica, Inc., 274 F.3d 577 (1st Cir. 2001); Snap-On Bus. Sols., Inc. v. O'Neil & Assocs., Inc., 708 F. Supp. 2d 669 (N.D. Ohio 2010).

[131] *See, e.g.*, *Power Ventures*, 252 F. Supp. 3d 765; QVC, Inc. v. Resultly, LLC, 99 F. Supp. 2d 525 (E.D. Pa. 2015); Craigslist, Inc. v. 3Taps, Inc., 942 F. Supp. 2d 962 (N.D. Cal. 2013).

[132] *See, e.g.*, *Tamburo*, 974 F. Supp. 2d 1199. In A.V. *ex rel*. Vanderhye v. iParadigms, LLC, the scraper in question was a commercial plagiarism detection software that was alleged to have scraped websites to build its comparative corpus, but the plaintiff in that particular case submitted material voluntarily, and the counterclaims at issue involved that plaintiff's use a third-party account to access the scraper's services. 562 F.3d 630 (4th Cir. 2009).

[133] VRCompliance LLC v. HomeAway, Inc., No. 1:11-CV-1088, 2011 WL 6779320, at *1 (E.D. Va. Dec. 27, 2011). The case was decided on procedural grounds. *Id*. at *5–6.

[134] United States v. Auernheimer, No. 11-cr-470, 2012 WL 5389142, at *1 (D.N.J. Oct. 26, 2012), *rev'd on other grounds*, 748 F.3d 525 (3d Cir. 2014) (scraping AT&T's website to reveal a data vulnerability and disclosing the fruits of this to an online publication). In the interest of full disclosure, I co-authored an *amicus* brief in this case in support of the defendant. *See* Brief for Digital Media Law Project as Amicus Curiae Supporting Defendant-Appellant, United States v. Auernheimer, 748 F.3d 525 (3d Cir. 2014) (No. 13-1816). While perhaps motivated by a sense of public interest, the defendant in question is as far from morally praiseworthy, a self-proclaimed neo-Nazi who operates one of the largest platforms for white supremacists on the Internet. Rachel Gutman, *Who Is Weev, and Why Did He Derail a Journalist's Career?*, ATLANTIC (Feb. 14, 2018), https://www.theatlantic.com/technology/archive/2018/02/who-is-weev/553295/ [https://perma.cc/LF3X-TZAH].

CFAA brought by a number of plaintiffs who engaged in scraping as part of their academic and journalistic pursuits.[135]

It is unclear what precisely accounts for this concentration of cases in the commercial arena. The text of the CFAA generally does not draw distinctions based on the purpose for which one accesses a computer without authorization.[136] Rather than any distinction in the statute, the prevalence of commercial suits may reflect a belief among website owners that commercially competitive scrapers are the only ones that cause harm worthy of the expense of a lawsuit. Web hosts might also be hesitant to pursue scrapers that have a public interest motivation, for fear of public backlash or unfavorable judicial precedent.

Claims against scrapers tend to be brought under the "obtaining information" provisions in 18 U.S.C. § 1030(a)(2)(C)[137] and the "computer fraud" provisions in § 1030(a)(4),[138] though a few also address the "damage" provisions in § 1030(a)(5).[139] Under the "obtaining information" provisions, one violates the CFAA when one "intentionally accesses a computer without authorization or exceeds authorized access, and thereby obtains [. . .] information from any protected computer."[140] One violates the "computer fraud" provision when one "knowingly and with intent to defraud, accesses a protected computer without authorization, or exceeds authorized access, and by means of such conduct furthers the intended fraud and obtains anything of value," with the caveat that the "thing of value" cannot be the use of the computer itself, unless such use is worth more than $5000 in a one-year period.[141]

Both provisions require a plaintiff or prosecutor to first show that a user accessed a computer "without authorization" or "exceed[ed] authorized access."[142] How precisely to interpret these phrases has been at the center of a

[135] Sandvig v. Sessions, No. 16-cv-1638 (JDB), 2018 WL 1568881 at *2 (D.D.C. Mar. 30, 2018).

[136] Though one could argue that an "intent to defraud," as required in 18 U.S.C. § 1030(a)(4) (2012), implies a certain degree of commerciality or at least pecuniary transfer that may not be met in some cases. *See* United States v. Czubinski, 106 F.3d 1069, 1074–75 (1st Cir. 1997). A commercial purpose can also escalate sentencing of crimes under §§ 1030(a)(2), (a)(3), and (a)(6). *See id*. § 1030(c)(2)(B)(i).

[137] It does not appear as though any case has applied the similar provisions in §§ 1030(a)(2)(A) and (B), which protect records from financial institutions and information from federal departments and agencies.

[138] *See, e.g*., Fidlar Techs., Inc. v. LPS Real Estate Data Sols., Inc., 810 F.3d 1075 (7th Cir. 2016).

[139] *See, e.g*., QVC, Inc. v. Resultly, LLC, 99 F. Supp. 3d 525 (E.D. Pa. 2015).

[140] § 1030(a)(2)(C). Under the most recent definition, a "protected computer" extends to any computer "which is used in or affecting interstate or foreign commerce." *Id*. § 1030(e)(2)(B).

[141] *Id*. § 1030(a)(4).

[142] The statute does not define "authorization," but defines "exceeds authorized access" recursively. *Id*. § 1030(e)(6) ("[T]he term 'exceeds authorized access' means to access a computer with authorization and to use such access to obtain or alter information in the

very large portion of the discussion about the CFAA[143] and is a main focus of this piece as well.

Scholars have endeavored to taxonomize the types of mechanisms that courts reference when analyzing questions of authorization. In a landmark work from 2003, Prof. Orin Kerr divided decisions addressing the CFAA's "authorization" question into three categories.[144] First, he identified decisions that looked to the "intended function" of the technology used to gain "access" to the computer in question — drawing principally from the 1991 Second Circuit CFAA case *United States v. Morris*, a case that concerned a computer science student who sent a self-replicating worm through exploits of protocols on the early Internet.[145] This test looks to find a particular technological tool that a defendant used to access a computer and then see whether the defendant used the tool either in accordance with its designed purpose or in a way that it is otherwise popularly employed.[146] As Kerr observes, the "intended function" test is a blended consideration of the computer's code-based mechanisms of access and the social norms surrounding the use thereof.[147]

Second, he identified cases that find a lack of authorization due to the misconduct of parties who may owe a duty to the computer's owner, such as an employee on a work computer.[148] As Kerr notes, this is a "strikingly broad" definition of unauthorized access, as it would find felonious conduct "whenever an employee uses a computer for reasons contrary to an employer's interest."[149] Third, Kerr identifies a series of cases finding that defendant's breach of an agreement governing their use of the computer in question rendered their access "unauthorized" under the CFAA.[150]

---

computer that the accesser is not entitled so to obtain or alter[.]"). The damage provisions under the CFAA use a slightly different phrasing, and punish those who access "without authorization" only, or "cause[] damage without authorization." *Id.* § 1030(a)(5).

[143] *See, e.g.*, Patricia L. Bellia, *A Code-Based Approach to Unauthorized Access Under the Computer Fraud and Abuse Act*, 84 GEO. WASH. L. REV. 1442 (2016); Kerr, *Norms*, *supra* note 116; Michael J. Madison, *Authority and Authors and Codes*, 84 GEO. WASH. L. REV. 1616 (2016); Matthew Gordon, *A Hybrid Approach to Analyzing Authorization in the Computer Fraud and Abuse Act*, 21 B.U. J. SCI. & TECH. L. 357 (2015); Cyrus Y. Chung, *The Computer Fraud and Abuse Act: How Computer Science Can Help with the Problem of Overbreadth*, 24 HARV. J. L. & TECH. 233 (2010).

[144] Kerr, *Cybercrime's Scope*, *supra* note 73, at 1628-32.

[145] *Id.* at 1629–30 (citing United States v. Morris, 928 F.2d 504 (2d Cir. 1991).

[146] *Id.* at 1630.

[147] *Id.*

[148] *Id.* at 1633

[149] *Id.*

[150] *Id.* at 1637–40. After laying out this framework, Kerr argues that only access circumventing code-based restrictions, such as the restrictions at issue in *Morris*, should be a valid basis for CFAA claims. *See Id.* at 1643-45. He supports this proposal with a number of arguments, including his assertion that "limiting the scope of computer misuse statutes to the circumvention of code-based restrictions would let criminal law advance two vitally im-

More recently, in a piece from 2016, Professor Patricia Bellia identified five different "interpretive paradigms" courts use to assess authorization under the CFAA.[151] First is the "agency paradigm," which is largely similar to Kerr's employee misconduct paradigm above.[152] Second is what Bellia calls the "norms-of-access paradigm," which she cites as the method adopted by the Second Circuit in the *Morris* decision, taking this decision out of the blended code- and norms-based category that Kerr placed the case in above, and casting the case instead as one where the "court developed a concept of authorized access based on its understanding of how one *ought* to use the technology in question"[153] Bellia breaks more contractual cases into two categories: a "policy paradigm" to encompass authorization based on terms of use and other unilateral statements by computer owners,[154] and a "contract paradigm" that looks more specifically at whether a fully-formed contract existed between the user and the computer owner.[155] Finally, Bellia notes that, while no never fully adopted in an appellate decision, some courts have suggested use of a purely code-based paradigm.[156]

As shown in the sections that follow, courts at different times have looked to different types of mechanisms in web scraping cases. For ease of discussion, I roughly categorize the restrictions in question as being code-based, contract-based, or based on a normative understanding.

A. The 2000s: Anything Can Inform Authorization

In the first decade of web scraping cases, courts embraced virtually all of the theories of authorization set out above. It seemed in this period that *any* mechanism could be used to determine that the scraper's access was unauthorized and therefore in violation of the statute. From the turn of the millennium[157] un-

portant and often conflicting goals of Internet regulation: first, to allow Internet users to enjoy as much freedom as possible to do as they wish online, and, second, to protect the privacy and security of Internet users and their data." *Id*. at 1649.

[151] Bellia, *supra* note 144, at 1445.

[152] *Id*. at 1446–47.

[153] *Id*. at 1449.

[154] *Id*. at 1451-55

[155] *Id*. at 1455-56.

[156] *Id*. at 1457–60. Bellia, like Kerr, has argued for a drive towards a code-based interpretation of authorization. *Id*. at 1476.

[157] The CFAA made an appearance in some of the earliest cases on the lawfulness of web scraping, including the now-famous eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058 (N.D. Cal. 2000). The court in that case focused its analysis solely on the trespass to chattels doctrine. *Bidder's Edge*'s usual analytic companion, Ticketmaster Corp. v. Tickets.com, Inc., also confined its scraping analysis to trespass to chattels doctrine. No. 99-CV-7654 HLH(BQRX), 2000 WL 525390 (C.D. Cal. Mar. 27, 2000). The first case to actually analyze CFAA liability appears to be Register.com, Inc. v. Verio, Inc., 126 F. Supp. 2d 238 (S.D.N.Y. 2000).

til courts began to shift their analysis in 2009,[158] there appear to be nine opinions that discuss liability under the CFAA for web scrapers, resulting in preliminary injunctions in a handful of cases and denied motions to dismiss in a couple of others.[159] While these cases leave some ambiguity based on their procedural posture, these early opinions appeared to suggest that virtually any signal of a website's displeasure about a scraper is sufficient to put the scraper on notice that subsequent access would be "unauthorized." Courts found, for example, that the violation of a restriction on the use of information could retroactively make the scraper's access unauthorized,[160] as could breaching a term of service,[161] or accessing a public website after receiving express warnings to stay away.[162] In one case, the court found that the filing of the complaint in the case itself served to signal that subsequent access was unauthorized, thus giving grounds for a preliminary injunction in the very same case.[163] Two cases from this period suggested that use of a third-party user account with permission of the account holder but without permission of the website could form

[158] I draw this line specifically at the decision of the non-scraping case LVRC Holdings LLC v. Brekka, 581 F.3d 1127 (9th Cir. 2009), for reasons discussed in Section III.B. below.

[159] The cases analyzed are, in chronological order, Register.com, Inc. v. Verio, Inc., 126 F. Supp. 2d 238 (S.D.N.Y. 2000), *aff'd on other grounds*, 356 F.3d 393 (2d Cir. 2004) (allowing preliminary injunction against scraper); EF Cultural Travel BV v. Explorica, Inc., 274 F.3d 577 (1st Cir. 2001) (allowing preliminary injunction against scraper); EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58 (1st Cir. 2003) (suggesting under the same facts as *Explorica, Inc*. that the scraper themselves would not be enjoined); Sw. Airlines Co. v. Farechase, Inc., 318 F. Supp. 2d 435 (N.D. Tex. 2004) (motion to dismiss brought by scraper denied); Facebook, Inc. v. ConnectU LLC, 489 F. Supp. 2d 1087 (N.D. Cal. 2007) (motion to dismiss for California CFAA equivalent brought by scraper denied); Healthcare Advocates, Inc. v. Harding, Early, Follmer & Frailey, 497 F. Supp. 2d 627 (E.D. Pa. 2007) (summary judgment in favor of scraper); Ticketmaster LLC v. RMG Techs., Inc., 507 F. Supp. 2d 1096 (C.D. Cal. 2007) (preliminary injunction issued against scraper, but not on CFAA grounds); A.V. v. iParadigms, LLC, 544 F. Supp. 2d 473 (E.D. Va. 2008), *rev'd sub nom*., A.V. ex rel. Vanderhye v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009) (in an unusual posture, counterclaim brought by the scraper against a party objecting to other aspects of scraper's activity initially denied on summary judgment, then reversed on appeal). While not indexed in either of the major online case databases, the District of Massachusetts decision that was appealed in the *Explorica, Inc*. and *Zefer Corp*. cases also granted a preliminary injunction against the scraper. *See Explorica Inc*., 274 F.3d at 580.

[160] *Id*. at 583–84; *Register.com*, 126 F. Supp. 2d at 251 ("[B]oth Verio's method of accessing the WHOIS data and Verio's end uses of the data violate the CFAA.").

[161] *Farechase*, 317 F. Supp. 2d at 439 (terms of use banned "any deep-link, page-scrape, robot, spider or other automatic device, program, algorithm or methodology which does the same things").

[162] *Id*. at 439-40.

[163] *Register.com*, 126 F. Supp. 2d at 249 ("[I]t is clear since at least the date this lawsuit was filed that Register.com does not consent to Verio's use of a search robot[.]").

the basis of access without authorization, but neither squarely addressed the claim.[164] This was, to put it simply, a very uncertain time for web scrapers.[165]

Three opinions sided in favor of the scraper during this time. One court denied a preliminary injunction on the grounds that the website host failed to factually substantiate its claims.[166] The other two based their decision on the absence of any authorization mechanism at all[167] — though each also suggested that a well-deployed terms of use notice on the website may have changed their analysis.[168] The would-be dissent from the late Judge Fred I. Parker in the Second Circuit's *Register.com, Inc. v. Verio, Inc.* decision also indicated that it would have found in favor of the scraper on a CFAA claim, as the website failed to show the requisite harm to bring a civil action.[169] Given the limited room by which the scraper escaped liability in each case, it is hard to find grounds to believe that these courts were making their decisions based on a narrow reading of the CFAA. Each case instead seemed one minor factor away from finding liability.

Perhaps the only positive indication for scrapers in this time came in the form of dicta in the First Circuit's decision in *EF Cultural Travel BV v. Zefer Corp.*, which suggested that there may come a point where public policy would prevent a court from finding CFAA liability based on contractual restrictions

[164] *iParadigms, LLC*, 544 F. Supp. 2d at 479, 486 (use of a username and password found on the Internet may be unauthorized, but court found a lack of sufficient loss to meet the civil action threshold, the Fourth Circuit reversing on this point but not addressing the merits); *ConnectU*, 489 F. Supp. 2d at 1091 (examining the California CFAA equivalent, and finding that even if use of a third-party account may be permitted access, the subsequent copying of information with authorization likely violated the statute).

[165] This state of affairs was referenced later in *Tamburo v. Dworkin*, when the court had to analyze whether a statement that data scraped off a website was "stolen" could be grounds for a defamation claim. The court found it was protected opinion in part because of how unsettled the law around web scraping was in 2004, when the statement was made. Tamburo v. Dworkin, 974 F. Supp. 2d 1199, 1215–16 (N.D. Ill. 2013).

[166] Ticketmaster LLC v. RMG Techs., Inc., 507 F. Supp. 2d 1096, 1113 (C.D. Cal. 2007). The court did, however, indicate that it would find that the scraper accessed the website without authorization, and enjoined the scraper on other grounds.

[167] EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58, 63 (1st Cir. 2003); Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey, 497 F. Supp. 2d 627, 649 (E.D. Pa. 2007). *Healthcare Advocates* is an especially unusual scraping case, as it concerns the archival copies of websites made by the Internet Archive which, as it happens, retroactively applies the Robots Exclusion Standard to previously-archived material. The court stressed that its analysis of access questions was closely tailored to these unique facts. *Id.* at 643.

[168] *Zefer*, 318 F.3d at 63 ("If EF wants to ban scrapers, let it say so on the webpage or a link clearly marked as containing restrictions"); *Healthcare Advocates*, 497 F. Supp. 2d at 649 (distinguishing *Farechase*, 318 F. Supp. 2d at 435, on the grounds that in that case "the defendant had agreed not to scrape the information," an agreement done through terms of service in that case).

[169] *Register.com, Inc.*, 356 F.3d at 439–40 (Opinion of Parker, J.). For the history of this opinion and its unusual format, see the discussion in *supra* note 125.

on a public website.[170] In support of this point, the First Circuit cited a pair of cases concerning undercover journalism, where prior courts had limited damages and rejected some claims based on trespass and breaches of duties on the part of the journalists posing as customers or low-level employees, out of a free speech concern.[171] This idea would resurface again in the most recent set of cases on web scraping,[172] but first courts would have to narrow in, and then expand out, interpretation of the CFAA through the intervening decade and a half.

B. The Early 2010s: A Narrower Reading, and a Lean Toward Technology

The next round of opinions concerning web scraping and the CFAA — starting in 2009 and continuing until 2013 — signaled a slight trend towards limiting the law's application. This follows a pattern in the CFAA cases more generally, where (to use popular terminology[173]) the earlier decisions that found "broad" reasons to find access without authorization under the CFAA[174] began to give way to courts adopting a "narrow" view on unauthorized access.[175]

To unpack this further, courts who narrow the CFAA appear to do so along two major lines: looking explicitly to technical controls instead of controls set by contract or principles of duty,[176] and policing against the application of mere "use restrictions" (as opposed to "access restrictions") to govern unauthorized access under the CFAA.[177] While other circuits have now joined this trend,[178] two Ninth Circuit opinions served as an early catalyst*: LVRC Hold-*

[170] 318 F.3d at 63.

[171] *Id*. at 63 (citing Food Lion, Inc. v. Capital Citites/ABC Inc., 194 F.3d 505 (4th Cir. 1999); Desnick v. ABC, 44 F.3d 1345, 1351 (7th Cir. 1995)).

[172] *See infra* Section III.D.

[173] *See, e.g.*, Mayer, *supra* note 123.

[174] *See, e.g.*, Int'l Airport Ctrs., LLC v. Citrin, 440 F.3d 418, 420 (7th Cir. 2006).

[175] *See, e.g.*, United States v. Nosal (*Nosal I*), 676 F.3d 854, 863 (9th Cir. 2012) (en banc).

[176] *See, e.g.*, Adv. Micro Devices, Inc. v. Feldstein, 951 F. Supp. 2d 212, 217 (D. Mass. 2013) (noting, and adopting, a narrow interpretation that "reflects a technological model of authorization, whereby the scope of authorized access is defined by the technologically implemented barriers that circumscribe that access"). The Fourth Circuit has also hinted at this interpretation. WEC Carolina Energy Sols. LLC v. Miller, 687 F.3d 199, 207 (4th Cir. 2012) (finding no unauthorized access when the defendant had access on a technical level).

[177] *Nosal I*, 676 F.3d at 863–64; Wentworth-Douglass Hosp. v. Young & Novis Prof. Ass'n, No. 10-cv-120-SM, 2012 WL 2522963, at *4 (D.N.H. June 29, 2012). This differentiation appears in some cases several years before the decisions discussed here. *See* Int'l Assoc. of Machinists & Aerospace Workers v. Werner-Masuda, 390 F. Supp. 2d 479, 499 (D. Md. 2005) ("[The CFAA does] not prohibit the unauthorized disclosure or use of information, but rather unauthorized access.").

[178] *See, e.g.*, United States v. Valle, 807 F.3d 508, 528 (2d Cir. 2015); *WEC*, 687 F.3d at 206.

ings LLC v. Brekka in 2009[179] and the *en banc* decision in *United States v. Nosal (Nosal I)* in 2012.[180] *Brekka* concerned a lawsuit brought by an employer against a former employee, Christopher Brekka, who, while still employed, sent emails from his work computer to a personal computer.[181] The employer argued that Brekka accessed his work computer without authorization because he sent the information on his work computer "to further his own personal interests, rather than the interests of LVRC."[182] The court rejected this duty-based theory of authorization, finding that because Brekka had permission to both use the computer and access the documents in question he did not violate the CFAA.[183]

In *Nosal I*, the *en banc* Ninth Circuit considered a similar set of facts involving an employee at an executive recruiting firm who, after he left, convinced his (then still-employed) former coworkers to send him customer information, which would allow him to launch a rival business.[184] The employer argued that the employees' access to the customer database was "unauthorized" because use of the database for anything other than official business violated their employment contract.[185] The court found that such "use" restrictions were improper grounds for liability under the CFAA.[186] Although the case did not involve access to public websites, the court suggested in dicta that its reasoning would also bar claims based on violations of restrictions memorialized in websites' terms of use.[187] Further, nearly all of the court's examples of actions that *would* be unauthorized access were more akin to what most would call "hacking," or circumvention of code-based controls to a computer, though the court did not go so far as to explicitly require this.[188]

---

[179] 581 F.3d 1127 (9th Cir. 2009).

[180] *Nosal I*, 676 F.3d 854.

[181] *Brekka*, 581 F.3d at 1129–30.

[182] *Id.* at 1132. The plaintiff claimed both that the access was without authorization and exceeded authorized access, and the court interpreted the two the same way. *Id.* at 1135 n.7.

[183] *Id.* at 1135.

[184] 676 F.3d at 856

[185] *Id.*

[186] *Id.* at 863–64.

[187] *Id.* at 860–61.

[188] *See id.* at 858 (noting that an employee who "circumvents the security measures" on a system would exceed authorized access); *id.* (suggesting that access "without authorization" would apply to "outside hackers," and "exceeds authorized access" would apply to "inside hackers"); *id.* at 863 (stating that the purpose of the CFAA is to punish "the circumvention of technological access barriers"); *but see id.* at 857 ("[A]ssume an employee is permitted to access only product information on the company's computer but accesses customer data: He would 'exceed[] authorized access if he looks at the customer lists."). The court also uses the example of an employee who uses another's login credentials, which could be argued as a technical control or not. *See id.* at 858. That said, courts in the Ninth Circuit following *Nosal I* were careful not to expressly adopt a strictly-code-based theory of authorization. *See, e.g.*, Weingand v. Harland Fin. Sols., Inc., No. C-11-3109 EMC, 2012 WL 2327660, at

This trend away from finding liability based on mere "use restrictions" and toward examination of technological forms of authorization should, at least in theory, allow for greater tolerance of web scraping. After all, as noted in Section II, a web scraper covers no more ground than a web browser itself, and so the technical access which allows a person to view a website, should likewise grant someone the ability to scrape the same files.[189] Furthermore, the most common form of contractual restriction on scraping, a website's terms of use, usually only impart "use restrictions" — that is, they allow you to access the site and merely place restrictions on what you can do with the information *after* you arrive.[190] Both the lean towards code-based mechanisms of authorization and the policing against finding CFAA liability based on "use" restrictions would be strong steps toward protection of web scrapers.

And indeed, scraping cases from this period seem to take some tentative steps toward this narrowing. There are twelve substantive opinions on web scraping and the CFAA issued between the decision in *Brekka* and the August 2013 opinion in *Craigslist Inc. v. 3Taps Inc.*, which changed how courts analyze scraping, as further discussed in Section III.C.[191] The period saw five cas-

*3 (N.D. Cal. June 19, 2012). For a fuller discussion of the interpretation of *Nosal I* and whether it allows an exclusively code-based claim of unauthorized access, see Mayer, *The "Narrow" Interpretation*, *supra* note 45.

[189] *See supra* notes 105–107 and accompanying text.

[190] Even if written in the form of a condition of access, courts have policed against attempts to convert use restrictions into access restrictions in order to make a claim under the CFAA. Wentworth-Douglass Hosp. v. Young & Novis Prof'l Ass'n, No. 10-CV-120-SM, 2012 WL 2522963, at *4 (D.N.H. June 29, 2012). Craigslist Inc. v. 3Taps Inc. makes this point as well, but then turns the argument on its head, as described in Section III.C. below. Craigslist Inc. v. 3Taps Inc, 964 F. Supp. 2d 1178, 1184–85 (N.D. Cal. 2013).

[191] They are, in chronological order, Snap-On Bus. Sols., Inc. v. O'Neil & Assocs., 706 F. Supp. 2d 669 (N.D. Ohio 2010) (scraper's summary judgment motion denied); Facebook, Inc. v. Power Ventures, Inc., No. C 08-05780 JW, 2010 WL 3291750 (N.D. Cal. July 20, 2010) (cross-motions for summary judgment denied); Oracle USA, Inc. v. Rimini St., Inc., No. 2:10-CV-00106-LRH-PAL, 2010 WL 3257933 (D. Nev. Aug. 13, 2010) (scraper's motion to dismiss denied); Cvent, Inc. v. Eventbrite, Inc., 739 F. Supp. 2d 927 (E.D. Va. 2010) (scraper's motion to dismiss granted); Snapt Inc. v. Ellipse Commc'ns Inc., No. 3:09-CV-0661-O, 2010 WL 11542004 (N.D. Tex. Sept. 28, 2010) (summary judgment for putative scraper); Facebook, Inc. v. Power Ventures, Inc., 844 F. Supp. 2d 1025 (N.D. Cal. 2012) (summary judgment against scraper); EarthCam, Inc. v. Oxblue Corp., No. 1:11-cv-02278-WSD, 2012 WL 12836518 (N.D. Ga. Mar. 26, 2012) (scraper's motion to dismiss denied in part); CollegeSource, Inc. v. AcademyOne, Inc., No. 10-3542, 2012 WL 5269213 (E.D. Pa. Oct. 25, 2012) (summary judgment for scraper); United States v. Auernheimer, No. 11-cr-470 (SDW), 2012 WL 5389142 (D.N.J. Oct. 26, 2012) (denying motion to dismiss indictment); Citizens Info. Assocs. v. Justmugshots.com, No. 1-12-CV-573-LY, 2013 WL 12076563 (W.D. Tex. Feb. 26, 2013) (scraper's motion to dismiss granted); Craigslist Inc. v. 3Taps Inc., 942 F. Supp. 2d 962 (N.D. Cal. 2013) (scraper's motion to dismiss denied); EarthCam, Inc. v. Oxblue Corp., No. 1:11-cv-2278-WSD, 2013 WL 11904713 (N.D. Ga. July 19, 2013) (scraper's motion to dismiss granted in relevant part). While not a scraping

es where claims against scrapers were either adjudicated in the scraper's favor or dismissed.[192] Two of the opinions rejected use of a website terms of service to inform a CFAA claim.[193] The opinions at first read a bit like those from the earlier period, as in both cases the courts focus on the lack of a validly-formed contract to make the terms binding. Each case, however, goes one step further, and emphasizes the public nature of the content in question as an additional reason for finding a lack of CFAA liability.[194] No similar argument can be found in the cases from the earlier period. (The three other cases from this period are decided on fact-specific grounds: failure to prove that a scraper actually accessed the host's computer,[195] failure to show the requisite level of loss to bring a civil action,[196] and failure to bring a timely claim.[197])

And even the cases that *do* find "unauthorized" access reflect a narrowed approach to the statute. For example, the 2010 decision by the Northern District of California in *Facebook, Inc. v. Power Ventures, Inc.* considered Power.com's use of third-party Facebook accounts (with the account holder's permission) to scrape user data in order to build a social media aggregation service.[198] While allowing a claim under California's CFAA analogue to proceed, the court found that use of Facebook terms of use to determine authorization would "create a constitutionally untenable situation in which criminal penalties could be meted out on the basis of violating vague or ambiguous terms of use." The court held that in order to prove its claim, Facebook needed

case, a similar discussion can be found in Koch Indus., Inc. v. Does, No. 2:10CV1275DAK, 2011 WL 1775765 at *8 (D. Utah May 9, 2011).

[192] *EarthCam, Inc.*, 2013 WL 11904713, at *5; *Citizens Info. Assocs.*, 2013 WL 12076563 at *4; *CollegeSource*, 2012 WL 5269213, at *23; *Snapt*, 2010 WL 11542004, at *6; *Cvent*, 739 F. Supp. 2d at 927. The decision in *Snapt* was also affirmed with little elaboration during this time, Snapt Inc. v. Ellipse Commc'ns. Inc., 430 F. App'x 346 (5th Cir. 2011), and the decision in *CollegeSource Inc.* was similarly affirmed a few years later. CollegeSource, Inc. v. AcademyOne, Inc., 597 F. App'x 116 (3d Cir. 2015).

[193] *CollegeSource,* 2012 WL 5269213, at *14 ("[B]ecause AcademyOne was under no obligation to abide by any terms of use as to the CataLink acess, it did not exceed authorized access."); *Cvent*, 739 F. Supp. 2d at 933 ("Cvent's website, including its CSN database, is therefore not protected in any meaningful fashion by its Terms of Use or otherwise.").

[194] *See CollegeSource*, 2012 WL 5269213, at *14 (citing *Brekka*, then noting that "the record does not reflect any evidence of a breach of security or 'hacking' by AcademyOne" and that the information in question "is available on the Internet and does not require a password or individualized access"); *Cvent Inc.*, 739 F. Supp. 2d at 933 (distinguishing this case from earlier cases in part because "the entire world was given unimpeded access to Cvent's website").

[195] *Snapt*, 2010 WL 11542004, at *2.

[196] *Citizens Info. Assocs.*, 2013 WL 12076563, at *3–4.

[197] *EarthCam*, 2013 WL 11904713, at *5. The case did allow other claims to proceed on separate grounds.

[198] The factual background of the case can be found at Facebook, Inc. v. Power Ventures, Inc., No. C 08-5780 JF (RS), 2009 WL 1299698, at *1–2 (N.D. Cal. May 11, 2009).

to show that Power Ventures accessed Facebook "in a manner that overcomes technical or code-based barriers."[199] Two years later, the court granted summary judgment for Facebook on this point, after the website showed that Power Ventures circumvented technical barriers by designing its system to deliberately evade Internet Protocol (IP) address blocks put in place by Facebook.[200] (On appeal the Ninth Circuit adopted a very different analysis, as discussed below.)[201]

In addition to this case, two others considered whether the scraper's use of a third-party's website account could form the basis of CFAA liability. As Kerr recently noted, characterizing the analysis of these third-party-account cases as code-based versus contract-based can be difficult, as they tend to include elements of both paradigms.[202] The court in *Power Ventures* focused on the code-based elements of the scraper's access, whereas the other two cases in this period focused instead on the contract between the third-party user and the platform. In one, the Northern District of Ohio denied a scraper's motion for summary judgment, finding that the authorization at issue turned on disputed terms in the agreement between the website host and its user.[203] In the other, an early decision in the long-running *Oracle USA, Inc. v. Rimini Street, Inc.* case, the court denied a motion to dismiss with little analysis, but specifically noted that the third-party user breached its license with the website by allowing the scraper to access the site.[204]

The one criminal scraping case from this period defied this narrowing trend, and explicitly rejected a code-based limitation to the CFAA, though the case was reversed on different grounds on appeal.[205] And while the case never re-

[199] Facebook, Inc. v. Power Ventures, Inc., No. C 08-05780 JW, 2010 WL 3291750, at *11 (N.D. Cal. July 20, 2010). This is also quite similar to the analysis initially undertaken in the court in *3Taps*, before pivoting to the expanded analysis discussed below. Craigslist Inc. v. 3Taps Inc., 942 F. Supp. 2d 962, 968–69 (N.D. Cal. 2013).

[200] Facebook, Inc. v. Power Ventures, Inc., 844 F. Supp. 2d 1025, 1038 (N.D. Cal. 2012), *aff'd in part and rev'd in part,* 844 F.3d 1058 (9th Cir. 2016).

[201] *See infra* notes 242–245 and accompanying text.

[202] *See* Kerr, *Norms*, *supra* note 116, at 1174.

[203] Snap-On Bus. Sols., Inc. v. O'Neil & Assocs., Inc., 708 F. Supp. 2d 669, 678 (N.D. Ohio 2010). The case resulted in a verdict against the scraper, though in a follow-on case involving the defendant's insurer, the owner of the account in question, a subsidiary of Mitsubishi, still maintained that it was the lawful owner of the data held by the website host and should have had authority to grant the scraper access in this way. Axis Surplus Ins. Co. v. Mitsubishi Caterpillar Forklift Am., Inc., No. H-11-3745, 2012 WL 1788171, at *4 (S.D. Tex. May 16, 2012).

[204] No. 2:10-CV-00106-LRH-PAL, 2010 WL 3257933, at *3 (D. Nev. Aug. 13, 2010). The Ninth Circuit ultimately found that the restrictions in Oracle's agreement with its users cannot form the basis of a claim under California or Nevada's CFAA equivalents. *See* Oracle USA, Inc. v. Rimini St., Inc., 879 F.3d 948, 962 (9th Cir. 2018).

[205] United States v. Auernheimer, No. 11-CR-470 (SDW), 2012 WL 5389142, at *3 n.1 (D.N.J. Oct. 26, 2012).

sulted in an opinion on point, the question of how technology should inform CFAA cases was a topic of debate following what was likely the most famous scraping case of this time, the prosecution of Aaron Swartz. Swartz was a Harvard researcher who scraped the contents of the JSTOR academic article database for a still-unknown project.[206] After Swartz's death, members of Congress introduced (and re-introduced) a bill entitled "Aaron's Law," which would categorically prohibit interpretations of CFAA authorization based on violations of terms of use.[207] Scholars at the time debated whether such a law would have actually helped Swartz in that case, or whether the technological controls at issue, such his evasion of IP and media access control address (MAC address) filters, could still be used to find liability.[208]

In sum, scrapers in this period still found themselves facing potential liability, but had new theories at their disposal to rebut such claims, such as arguments that a mechanism in question was a mere "use restriction," or that the mechanisms setting authorization should be more code-based to have legal effect. Subsequent cases could have then turned to the finer questions around the CFAA as applied to common web design mechanisms and controls, and looked in detail at user accounts, IP address blocking, or MAC address filters. But instead, courts picked up a new concept for interpreting authorization, and in so doing, brought their decisions closer in line to the first decade of web scraping litigation, where nearly any mechanism could be used to demonstrate that a scraper's access was "unauthorized."

C. The Mid 2010s: Brekka's "Revocation" Backdoor

The narrowing trend of the early 2010s was cut short soon after its adoption. More recently, courts seized upon some of the extraneous language in *Brekka*,[209] and used it to turn the analysis of that case inside out. Instead of focusing

[206] I analyzed the CFAA application to the case at some length shortly after his death. Andy Sellars, *The Impact of "Aaron's Law" on Aaron Swartz's Case*, DIGITAL MEDIA L. PROJECT (Jan. 18, 2013), http://www.dmlp.org/blog/2013/impact-aarons-law-aaron-swartzs-case [https://perma.cc/9NMF-R6RY]. With the benefit of hindsight, the emphasis I placed on code-based barriers setting authorization after *Nosal I* has not been as strong as I thought it would be at the time, given the cases discussed in Section III.C.

[207] H.R. 1918, 114th Cong. (2015); H.R. 2454, 113th Cong. (2013).

[208] *See* Sellars, *supra* note 206.

[209] While the language in question is prominent, there is good reason to consider it dicta. It certainly is not dispositive; under the specific holding in *Brekka*, the court found that the defendant had authorization to access the computer during the relevant time period, and the facts did not show that he accessed it after his employment terminated. LVRC Holdings LLC v. Brekka, 581 F.3d 1127, 1135–36 (9th Cir. 2009). At several times throughout the opinion, though, the Ninth Circuit opined as to whether such subsequent access would have violated the CFAA, if shown. *See id*. at 1135 ("Rather, we hold that a person uses a computer 'without authorization' under §§ 1030(a)(2) and (4) when the person has not received permission to use the computer for any purpose . . . or when the employer has rescinded permission to access the computer and the defendant uses the computer anyway."); *see also*

on the distinction between "access" and "use" described above, these cases instead seize upon dicta in which the Ninth Circuit speculates about what might be "unauthorized access" under a different set of facts: when the computer owner "*has rescinded permission to access the computer and the defendant uses the computer anyway*."[210]

A series of cases, including some web scraping cases, focused on this "revocation" theory to develop two different heuristics for evaluating authorization, depending on whether one accesses the computer "without authorization" or whether one "exceeds authorized access."[211] This trend began with the scraping case *Craigslist, Inc. v. 3Taps Inc.*,[212] and reached its apex with a pair of Ninth Circuit cases from 2016: a revisit to *United States v. Nosal* (*Nosal II*) and the *Power Ventures* cases discussed above. The cases pull the CFAA doctrine in two distinct directions. On the one hand, when examining cases brought under a claim that the defendant "exceed[ed] authorized access" they keep the narrowing and technologically-leaning construction from *Brekka* and *Nosal I*. But at the same time, they radically broaden what could constitute access "without authorization," covering any situation in which a mechanism signaling that there was a revocation of access can be identified — be it technological or not, use-based or access-based.[213] Courts in this period also found defendants to have exceeded authorized access where "revocation" signals conflicted with other authorization mechanisms, such as a cease-and-desist letters sent to stop someone from accessing a website freely available as a technical matter.[214]

id. at 1136 ("There is no dispute that if Brekka accessed LVRC's information on the LOAD website after he left the company in September 2003, Brekka would have access the protected computer "without authorization" for purposes of the CFAA.").

[210] *Id*. at 1135 (emphasis added).

[211] United States v. Nosal (*Nosal II*), 844 F.3d 1024, 1035–36 (9th Cir. 2016) (citing the dicta from *Brekka*, and confining cases like *Nosal I* as cases concerning the "exceeds authorized access" prong exclusively); Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1066-67 (9th Cir. 2016) (finding a total bar to access to an essentially-public website "[o]nce permission has been revoked," citing *Brekka*, and limiting *Nosal I* to merely saying that terms of use violations, alone, cannot form the basis of liability). This is in some ways the inverse to the First Circuit's approach in its first web scraping case, which seemed to signal that it may confine the more technically-leaning "intended function" test from United States v. Morris, 928 F.2d 504 (2d Cir. 1991), to cases where a person accessed "without authorization" only. EF Cultural Travel BV v. Explorica, Inc., 274 F.3d 577, 582 n.10 (1st Cir. 2001).

[212] Craigslist Inc. v. 3Taps Inc., 964 F. Supp. 2d 1178, 1185 (N.D. Cal. 2013).

[213] *Nosal II*, 844 F.3d at 1033; *Power Ventures*, 844 F.3d at 1066–67.

[214] Eriq Gardner, *Can Hillary Clinton Be Barred from Visiting DonaldJTrump.com?*, HOLLYWOOD REP. (July 13, 2016), https://www.hollywoodreporter.com/thr-esq/can-hillary-clinton-be-barred-910326 [https://perma.cc/9B46-NSL2] (discussing the potential extended limits of *Power Ventures*). Interestingly, the court in *Power Ventures* explicitly retreated from a purely code-based version of authorization when considering revocation of access,

One can trace this change based solely on the citations to *Brekka* itself. Prior to the pivot in the *3Taps* opinion, two scraping cases cited *Brekka* to argue for a limitation in CFAA liability,[215] and three cases cited *Brekka* for unrelated reasons.[216] Of *3Taps* and the cases that followed, no case cites *Brekka* to argue for a narrow reading. Instead, five opinions cite *Brekka* to support a broad "revocation" theory,[217] and one cites *Brekka* for an unrelated reason.[218]

Interestingly, *3Taps* itself adopted this reworked interpretation of *Brekka* only after an earlier opinion in the same case cast doubt on use of the CFAA to challenge access to "information generally available" on a public website.[219] *3Taps* involved claims by the online classified ads website Craigslist against a series of services that helped aggregate and visualize Craigslist listings.[220] To prove that the services' access was unauthorized, Craigslist pointed to its terms of use, cease-and-desist letters it sent to the services, and IP blocks it imposed.[221] The court's initial opinion reads much like those from the CFAA era discussed in Section III.B. above: it rejected use of terms of service to govern authorization, citing *Nosal I* and noting that the terms of use contained "only 'use' restrictions, not true 'access' restrictions,"[222] and allowed the claim to

finding that IP blocking, alone, should not form the basis of such a finding. *Power Ventures*, 844 F.3d at 1068 n.5.

[215] Cvent, Inc. v. Eventbrite, Inc., 739 F. Supp. 2d 927, 933 (E.D. Va. 2010) (citing *Brekka* as supporting the district's slightly-earlier decision to limit claims based on use restrictions instead of access restrictions); Facebook, Inc. v. Power Ventures, Inc., No. C 08-05780 JW, 2010 WL 3291750, at *9 (N.D. Cal. July 20, 2010) (citing *Brekka* to specifically to find that "more recent CFAA cases militate for an interpretation [of the California CFAA equivalent] that does not premise permission to access or use a computer or computer network on a violation of terms of use").

[216] United States v. Auernheimer, No. 11-cr-470 (SDW), 2012 WL 5389142, at *2 (D.N.J. Oct. 26, 2012) (citing *Brekka* to note that courts usually use the "ordinary meaning" of terms "without authorization"); CollegeSource, Inc. v. AcademyOne, Inc., No. 10-3542, 2012 WL 5269213 at *13 (E.D. Pa. Oct. 25, 2012) (citing *Brekka* for the general definition of "exceeds authorized access"); Snap-On Bus. Sols., Inc. v. O'Neil & Assocs., Inc., 708 F. Supp. 2d 669, 677 (N.D. Ohio 2010) (noting generally that *Brekka* supports the "narrow definition" of unauthorized access).

[217] Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1066–67 (9th Cir. 2016); CouponCabin LLC v. Savings.com, Inc., No. 2:14-CV-39-TLS, 2017 WL 83337, at *3 (N.D. Ind. Jan. 10, 2017); CouponCabin LLC v. Savings.com, Inc., No. 2:14-CV-39-TLS, 2016 WL 3181826, at *4 (N.D. Ind. June 8, 2016); QVC, Inc. v. Resultly, LLC, 159 F. Supp. 3d 576, 595 (E.D. Pa. 2016); *3Taps*, 964 F. Supp. 2d at 1183.

[218] Oracle USA, Inc. v. Rimini St., Inc., 191 F. Supp. 3d 1134, 1142, *rev'd*, 879 F.3d 948 (9th Cir. 2018) (stating that courts construe the computer access statutes strictly).

[219] Craigslist Inc. v. 3Taps, Inc., 942 F. Supp. 2d 962, 969–70 (N.D. Cal. 2013).

[220] *Id*. at 966.

[221] *Id*. at 969.

[222] The court also found that web crawlers were not explicitly prohibited under the terms of use. *See id*. n.6.

proceed based on the IP block instead.[223] In passing, however, the court suggested that it might even find an even more narrow application of the CFAA to web scraping, raising what it called "a threshold question of whether the CFAA applies where the owner of an otherwise publicly accessible website takes steps to restrict access by specific entities, such as the owner's competitors."[224]

After the court accepted additional briefing on the issue and revisited the question, it changed its tune considerably.[225] The court answered its "threshold question" emphatically in favor of the CFAA's application to all of these circumstances, citing *Brekka* to observe that "computer owners have the power to revoke the authorizations they grant,"[226] and found that Craigslist "affirmatively communicated its decision to revoke" through its letter and IP-address blocks.[227] The court rejected concerns raised by the scrapers, citing *Nosal I*, that applying the CFAA to "use policies 'that most people are only dimly aware of and virtually no one reads or understands' . . . presents serious notice concerns and also threatens to 'transform whole categories of otherwise innocuous behavior into federal crimes.'"[228] The court found no notice concern here, because "[t]he notice issue becomes limited to how clearly the website owner communicates the banning," and in this case "Craigslist affirmatively communicated its decision" through these mechanisms.[229] That Craigslist only attempted to block access to the site because it disagreed with how someone was using otherwise-lawfully obtained data was irrelevant; all that mattered was that Craigslist sought to ban one particular visitor to its public website, and that visitor was on notice of that ban.[230]

Through this recasting of the CFAA analysis, the court once again opened the door to a wide array of authorization mechanisms that previously had been narrowed away. Six out of a total of fifteen substantive opinions from this period deny scrapers' motions to dismiss.[231] Rather than looking to code-based

[223] *See id*. at 969–70. The court does foreshadow the case's later focus on the cease-and-desist letter, which the court specifically notes in passing prohibited access to the site "for any purposes." *Id*. at 969.

[224] *Id*. n.8.

[225] Craigslist Inc. v. 3Taps Inc., 964 F. Supp. 2d 1178 (N.D. Cal. 2013).

[226] *Id*. at 1183 (citing LVRC Holdings LLC v. Brekka, 581 F.3d 1127 (9th Cir. 2009)).

[227] *Id*. at 1184.

[228] *Id*. (quoting United States v. Nosal (*Nosal I*), 676 F.3d 854, 860–61 (9th Cir. 2012)).

[229] *Id*.

[230] *See id*. at 1185 ("That [Craigslist] did so because of how 3Taps used Craigslist's information is true, but beside the point").

[231] The full list of cases are, in chronological order, *Id*. (denying scraper's motion to dismiss based on revocation of access); Fidlar Techs. v. LPS Real Estate Data Sols., No. 13-cv-4021, 2013 WL 5973938 (C.D. Ill. Nov. 8, 2013) (denying motion to dismiss); EarthCam, Inc. v. Oxblue Corp., 49 F. Supp. 3d 1210 (N.D. Ga. 2014) (summary judgment in favor of scraper); CollegeSource, Inc. v. AcademyOne, Inc., 597 Fed. App'x 116 (3d Cir. 2015) (affirming summary judgment in favor of scraper); Fidlar Techs. v. LPS Real Estate Data Sols.,

mechanisms of authorization or carefully parsing "access restrictions" versus "use restrictions," courts found CFAA violations based on mechanisms like a website telling the scraper their access is revoked,[232] or even the contents of a terms-of-service agreement.[233] With the focus placed on "revocation," questions about the legal impact of technical controls like user accounts or IP and MAC address filtering all fell away in favor of an analysis which asked whether the website owner used a technical control to signal a revocation of access, and whether the user understood this signal.[234] One case went as far as to find that the act of an IP-address block alone, without any additional communication, served as effective notice of revocation of access.[235] Such an analysis not only overlooks the finer questions around whether circumvention of an IP address block should be grounds for a federal felony, it puts every Internet user who ever confronted a "down" website in a curious moment of legal risk. The

82 F. Supp. 3d 844 (C.D. Ill. 2015) (summary judgment in favor of scraper); QVC, Inc. v. Resultly, LLC, 99 F. Supp. 3d 525 (E.D. Pa. 2015) (denying preliminary injunction against scraper); Fidlar Techs. v. LPS Real Estate Data Solutions, 810 F.3d 1075 (7th Cir. 2016) (affirming summary judgment in favor of scraper); QVC, Inc. v. Resultly, LLC, 159 F. Supp. 3d 576 (E.D. Pa. 2016) (denying scraper's motion to dismiss, noting revocation of access); CouponCabin LLC v. Savings.com, Inc., No. 14-cv-39, 2016 WL 3181826 (N.D. Ind. June 8, 2016) (same); Oracle USA, Inc. v. Rimini St., Inc., 191 F. Supp. 3d 1134 (D. Nev. 2016), *rev'd in relevant part* 879 F.3d 948 (9th Cir. 2018) (declining scraper's motion for judgment as a matter of law, based on a revocation theory); Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058 (9th Cir. 2016) (affirming summary judgment against scraper based on a revocation theory); CouponCabin LLC v. Savings.com, Inc., No. 14-cv-39, 2017 WL 83337 (N.D. Ind. Jan. 10, 2017) (denying scraper's motion for judgment on the pleadings, based on revocation of access); Facebook, Inc. v. Power Ventures, Inc., 252 F. Supp. 3d 765 (N.D. Cal. 2017) (re-affirming Ninth Circuit result against scraper, and determining damages); EarthCam, Inc. v. OxBlue Corp., 703 Fed. App'x 803 (11th Cir. 2017) (affirming summary judgment in favor of scraper); DHI Group, Inc. v. Kent, No. 16-cv-1670, 2017 WL 4837730 (S.D. Tex. Oct. 26, 2017) (denying scraper's motion to dismiss).

[232] *See CouponCabin LLC*, 2016 WL 3181826, at *4. The case notes that the plaintiff also tried to impose security measures to block access. *See id*. at *1.

[233] *See DHI Grp*., 2017 WL 4837730, at *4; *QVC*, 159 F. Supp. 3d at 597 ("[J]ust as a cease-and-desist letter would put a publisher on notice that its actions were prohibited, VigLink's Terms of Service . . . put Resultly on notice that QVC prohibited web-crawling"). Some decisions adopt non-revocation reasons for denying a motion to dismiss. *See, e.g*., *Fidlar Techs*., 2013 WL 5973938, at *7 n.7 (rejecting an argument that plaintiff failed to show adequate loss for a civil CFAA claim).

[234] *See, e.g*., Craigslist Inc. v. 3Taps Inc., 964 F. Supp. 2d 1178, 1184 (N.D. Cal. 2013) ("Craigslist affirmatively communicated its decision to revoke 3Taps' access through its . . . IP blocking efforts."); *id*. at 1186 n.7 ("IP blocking . . . is a real barrier, and a clear signal from the computer owner to the person").

[235] *See CouponCabin LLC*, 2017 WL 83337, at *3 (as to one party who did not receive a direct communication, "[r]evocation of website access would have been sufficient to give the Defendants constructive notice that they were without authorization to act as they allegedly did").

website might be down due to a server error or technical bug with the user's ISP. It might instead be due to a site-imposed block — in which case the next attempt to load the page is grounds for CFAA liability.[236] The "refresh" button was never meant to hold such legal weight.

The extremity of this approach is perhaps best illustrated in the Eastern District of Pennsylvania's opinion in *QVC, Inc. v. Resultly, LLC*.[237] The court in that case denied a motion to dismiss on a CFAA claim because the scraper Resultly had accepted a terms-of-service agreement with a third-party promotional-services partner named VigLink, which required Resultly to "comply with all rules, regulations and guidelines, as well as any applicable . . . terms and conditions and policies" provided by merchants affiliated with VigLink's service.[238] VigLink in turn had entered into two agreements with the merchant QVC, which placed limits on what sources of information VigLink could use in its operations. [239] The court found that QVC's restriction on sources of information would extend to web-crawling, and let the chain of contracts back to Resultly inform a CFAA claim against Resultly by QVC.[240] It takes the contractual due diligence of an M&A attorney to sort out potential liability under this framework.

Three opinions in two cases went beyond early-stage litigation and actually assigned or upheld liability on this revocation theory — though one case was later reversed on appeal, as discussed in Section III.D. below.[241] The most noteworthy of this group is the appeal in *Facebook, Inc. v. Power Ventures, Inc*. While the lower court decision offered an example of a more technologically leaning CFAA analysis, on appeal the Ninth Circuit dispensed with technical issues and instead found CFAA liability based on Power Ventures continued access after having received a cease-and-desist letter from Facebook.[242] The court also found a curiously mixed role for Facebook's terms of use. On the one hand, the court noted that Power Ventures was not subject to the terms of use as it was not itself a Facebook user.[243] But at the same time, the court seemed to approve of the substance of the cease-and-desist letter, which revoked access *because* Power allegedly breached Facebook's terms, a contract

[236] The CFAA punishes attempted access without authorization as well. 18 U.S.C. § 1030(b).

[237] 159 F. Supp. 3d 576 (E.D. Pa. 2016).

[238] *Id*. at 584–85.

[239] *Id*. at 581–82.

[240] *Id*. at 597.

[241] Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058 (9th Cir. 2016), *on remand*, 252 F. Supp. 3d 765 (N.D. Cal. 2017); Oracle USA, Inc. v. Rimini St., Inc., 191 F. Supp. 2d 1134 (D. Nev. 2016), *rev'd*, 879 F.3d 948 (9th Cir. 2018).

[242] *Power Ventures*, 844 F.3d at 1068 n.5 ("[B]ypassing an IP address, without more, would not constitute unauthorized use").

[243] *See id*. at 1069 ("Facebook and Power had no direct relationship, and it does not appear that Power was subject to any contractual terms that it could have breached").

to which they were not a party.[244] The court squares these two findings by noting that "in addition to asserting a violation of Facebook's terms of use, the cease and desist letter warned Power that it may have violated federal and state law."[245] But this just completes a logical circle, creating the unusual situation where a declaration that the law was broken made it so.

Five opinions in three cases found in favor of the scraper during this time period, though their analyses also feel a bit more like those from the first decade of web-scraping litigation.[246] One opinion found for the scraper on the grounds that the parties had not formed a valid contract over access to this particular data.[247] A second looked to the contract between the parties to analyze the CFAA claim and found that the contract did not prohibit the activity in question.[248] One case involved the use of a third-party account with the account holder's permission, which as noted above, always presents a somewhat hybrid technological and contractual analysis.[249] Here, though, the analysis followed the contract to find that the website by its terms did not prohibit sharing in this way, and also cast doubt on whether the defendant was even aware of the contract in the first place.[250] The Eleventh Circuit affirmed this decision on the grounds that the evidence was "too attenuated" to show that the scraper knew of the contract in question.[251] Other scrapers who faced liability under the "fraud" provisions of § 1030(a)(4) or the "intentional damage" provisions of § 1030(a)(5)(A) prevailed on the plaintiff's failure to substantiate the additional elements required in those cases.[252]

If the story ended here, it would be hard to suggest that scrapers today are on better legal footing than they were in the early 2000s. Indeed, the situation

---

[244] *See id*. at 1067.

[245] *Id*. n.3.

[246] CollegeSource, Inc. v. AcademyOne, Inc., 597 Fed. App'x. 116 (3d Cir. 2015); Fidlar Techs. v. LPS Real Estate Data Sols., Inc., 82 F. Supp. 3d 844 (C.D. Ill. 2015), *aff'd*, 810 F.3d 1075 (7th Cir. 2016); EarthCam, Inc. v. OxBlue Corp., 49 F. Supp. 3d 1210 (N.D. Ga. 2014), *aff'd*, 703 Fed. App'x. 803 (11th Cir. 2017).

[247] *See CollegeSource*, 597 Fed. App'x. at 130 ("[Defendant] obtained the materials in question without breaching any technological barrier or contractual term of use"). The court also refers to code-based mechanisms of authorization in some of its discussion. *See id*. at 129 (noting that defendants were not shown to have "hacked into technologically sequestered portions of the database").

[248] *See Fidlar*, 810 F.3d at 1082 ("We see no reason why LPS should have inferred that it could not download records through a completely different program that it designed. LPS's access to records was tied to the individual agreements with each county—agreements that did not require LPS to use the Laredo client and that Fidlar was not even party to").

[249] *See supra* note 202–204 and accompanying text.

[250] *See EarthCam*, 49 F. Supp. 3d at 1232.

[251] EarthCam, Inc. v. OxBlue Corp., 703 Fed. App'x 803, 809–10 (11th Cir. 2017).

[252] *See, e.g., Fidlar*, 82 F. Supp. 3d at 845 (no intent to defraud); QVC, Inc. v. Resultly, LLC, 99 F. Supp. 3d 525, 542-43 (E.D. Pa. 2015) (preliminary injunction order denied under § 1030(a)(5)(A) claim, as the court doubted scraper intended to take down website).

may have even deteriorated, as a website need only preface a lawsuit with a cease-and-desist letter to lay the grounds for a CFAA claim; should the user access the website again, regardless of the grounds for objection, the nature of the scraper's access, or how public the website was, they will have violated the CFAA.[253] Most recently, however, three courts have pulled away from the broad, revocation-based theory of CFAA liability, presenting a possible framework for future interpretations of the statute in web scraping cases that is more sensitive to both the technical similarities between web scraping and web browsing and the odd result of legally banning one user from a website that all others are allowed to access.

D. Today: Revisiting Application of the CFAA to Public Websites

In the past few months a trio of opinions have presented a contrary framework to the broad, revocation-based theory that has risen over the past few years. The first case to do so was *hiQ Labs, Inc. v. LinkedIn Corp.*, which the Northern District of California decided in August of last year and is presently on appeal to the Ninth Circuit.[254] The facts of the case bear a striking similarity to those in *3Taps*.[255] The popular business-oriented social network LinkedIn sent a cease-and-desist letter and imposed an IP block against a scraper who had used publicly-facing LinkedIn information to offer business analytics to its customers.[256] The scraper in turn brought a declaratory judgment action against the website, seeking a declaration that it did not violate the CFAA, and the court did in fact enjoin LinkedIn from "preventing hiQ's access, copying, or use of public profiles on LinkedIn's website" while the case was pending.[257] While scholars quickly noted this atypical result, it is far from the first case to raise concerns about application of the CFAA to public websites.[258] Nor was

[253] One of the only upper limits noted by the courts at this time comes from the Ninth Circuit's decision in *Power Ventures*, which suggested that that a system where "an automatic boilerplate revocation follows a violation of a website's terms of use" might be too close to a pure use restriction to be consistent with *Nosal I*. *See* Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067 n.1 (9th Cir. 2016).

[254] 273 F. Supp. 3d 1099 (N.D. Cal. 2017), *appeal filed* No. 17-16783 (9th Cir. filed Sept. 6, 2017).

[255] *See supra* note 220 and accompanying text. 3Taps appears to think so too, as it has now filed a declaratory judgment action against LinkedIn seeking a similar injunction against LinkedIn blocking its scraping. That case has now been stayed pending the outcome of the *hiQ* case. Stipulation and Order, Dkt. No. 10, 3Taps, Inc. v. LinkedIn Corp., No. 18-cv-855 (N.D. Cal. filed Mar. 7, 2018).

[256] *hiQ Labs*, 273 F. Supp. 3d at 1104.

[257] *Id.* at 1120.

[258] *See, e.g.*, CollegeSource, Inc. v. AcademyOne, Inc., No. 10-3542, 2012 WL 5269213, at *14 (E.D. Pa. Oct. 25, 2012) ("[B]ecause the documents [at] issue were available to the general public, AcademyOne did not access them without authorization."); Sw. Airlines Co. v. BoardFirst L.L.C., No. 3:06-CV-0891-B, 2007 WL 4823761, at *13–14 (N.D. Tex. Sept. 12, 2007) (in a non-scraping case, discussing the disputed authority around using terms of

this the first case to enjoin a website against blocking access to a scraper. A month before, the Western District of Wisconsin had issued a similar injunction in an antitrust case involving a software scraper, though that injunction was vacated on appeal.[259]

In rejecting LinkedIn's claim of unauthorized access, the *HiQ Labs* court distinguished the Ninth Circuit's decision in *Power Ventures* on the grounds that the data in that case was not "public," as one can only access Facebook content with a username and password.[260] The court cited in passing the *3Taps* case — which suggested the possibility of CFAA liability for accessing a public website — but signaled disagreement with that result, saying that "whether 'access' to a publicly viewable site may be deemed 'without authorization' under the CFAA where the website hosts purports to revoke permission is not free from ambiguity."[261] The court went on to cite *Nosal I* to support the notion that Congress, in passing the CFAA, meant to embrace hacking in the more traditional sense, rather than more ambiguous forms of unauthorized access.[262] The court supported this interpretation with policy considerations, noting that assigning CFAA liability when someone accesses a website in contravention of a written instruction would "effectuat[e] the digital equivalence of Medusa," and would leave open the possibility that website owners could block users for discriminatory, anticompetitive, or other improper reasons.[263] This is an echo of the same concern raised by the First Circuit in the early scraping case *EF Cultural Travel BV v. Zefer Corp*.[264] Finally, the court also cited an influential article by Orin Kerr, which argued for an interpretation of the CFAA that recognizes the "inherent openness of the web."[265]

In the second case, decided earlier this year, the Ninth Circuit signaled a possible change in its thinking around scraping with its decision in *Oracle USA, Inc. v. Rimini Street, Inc*.[266] The case concerned a software support service that scraped a website containing manuals and technical material for the

service on publicly-accessible websites to impute unauthorized access under the CFAA); Ticketmaster Corp. v. Tickets.com, Inc., No. CV 99-7654 HLH(BQRX), 2000 WL 525390, at *4 (C.D. Cal. Mar. 27, 2000) (noting in a trespass to chattels case that "it is hard to see how entering a publicly available web site could be called a trespass, since all are invited to enter.").

[259] Authenticom, Inc. v. CDK Global, LLC, No. 17-cv-318-jdp, 2017 WL 3017048 (W.D. Wisc. July 14, 2017), *vacated*, 874 F.3d 1019 (7th Cir. 2017).

[260] *hiQ Labs*, 273 F. Supp. 3d at 1109.

[261] *Id*.

[262] *Id*. at 1109–10.

[263] *Id*. at 1110–11.

[264] *See supra* notes 171–172 and accompanying text.

[265] *hiQ Labs*, 273 F. Supp. 3d at 1111 (citing Kerr, *Norms*, *supra* note 116, at 1162). The court in *Power Ventures* similarly cites this article, but proceeds to then find that this presumption of openness can be revoked. *See* Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067 n.2 (9th Cir. 2016).

[266] 879 F.3d 948 (9th Cir. 2018).

major enterprise software provider Oracle, so that the service could provide more effective assistance to Oracle users.[267] Oracle brought a claim under the California and Nevada equivalents to the CFAA, alleging that Rimini Street scraped the website in violation of its terms of service.[268] The Ninth Circuit held that the scraper had at least some level of access to the public website, and therefore could not have accessed the site "without authorization," as the scraper did in *Power Ventures*.[269] Nor did the scraper "exceed authorized access," as the court found that the terms only limited the "method" of accessing information, instead of limiting the access itself.[270]

The Ninth Circuit characterized its decision as consistent with both *Power Ventures* and its earlier opinions in the *Oracle* litigation, and on first blush, this result appears to be a consistent with the cases discussed in Section III.C above.[271] The opinion, however, neglected to mention a key fact that would seem to put the case in tension with these decisions — specifically, that Oracle allegedly told Rimini Street to stop scraping, and blocked its IP to prevent further access, which Rimini Street then circumvented.[272] This is precisely the sort of behavior that the "revocation" line of cases, including *Power Ventures*, found to generate liability.[273] The *Oracle* court did not mention either the IP blocks or the communication from Oracle, and instead focused its analysis on the text of the terms of use itself, although it added a slight hedge in its language by noting that Rimini Street had authorization "at least at the time when it took the data in the first instance."[274] Unless the Ninth Circuit meant to say that once a scraper begins to scrape a website with authorization it can complete the process even if it receives an objection — which would be a novel theory in the CFAA caselaw — that statement seems to be in tension with the

[267] *Id*. at 952.

[268] *Id*.

[269] *Id*. at 962.

[270] *Id*.

[271] *Id*. (citing Oracle USA, Inc. v. Rimini St., Inc., 191 F. Supp. 3d 1134, at 1139–40 (D. Nev. 2016)).

[272] *Oracle*, 191 F. Supp. 3d at 1140.

[273] Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067 (9th Cir. 2016) ("Facebook expressly rescinded . . . permission when Facebook issued its written cease and desist letter to Power[.]"); CouponCabin LLC v. Savings.com, Inc., No. 2:14-CV-39-TLS, 2016 WL 3181826, at *4 (N.D. Ind. June 8, 2016) ("By alleging that the Defendants knowingly and intentionally circumvented the Plaintiff's security measures after the Plaintiff blocked access . . . and communicated with the Defendants by demanding they cease and desist scraping-related activities, the Plaintiff has pled enough facts to survive dismissal."); Craigslist Inc. v. 3Taps Inc., 964 F. Supp. 2d 1178, 1184 (N.D. Cal. 2013) ("Craigslist affirmatively communicated its decision to revoke 3Taps' access through its cease-and-desist letter and IP blocking efforts.").

[274] *Oracle*, 879 F.3d at 962.

facts below.[275] The court may have meant instead to suggest that a letter or technical block effectuating a revocation must be based on a legitimate, extraneous legal reason to hold any weight, which would explain its focus on the underlying contractual terms. This, however, would also seem to be in tension with the approach taken in *3Taps* and *Power Ventures*, which did not consider why the websites revoked access to their servers.[276]

The third case of this period was also the first to bring serious constitutional challenges to applications of the CFAA that prohibit web scraping: the District court for the District of Columbia's case *Sandvig v. Sessions*.[277] *Sandvig* was brought by a group of scholars and journalists who used web scraping and other technical tools as part of their research.[278] They were aware that the terms of service on many of the platforms they studied banned the techniques that they had used, including scraping, use of "sock puppet" accounts, and other common computational social science techniques.[279] The researchers argued that the First Amendment should bar enforcement of the CFAA based on such violations, as it implicated their rights to record or preserve information, and to publish the information that they found.[280] The court agreed that scraping "plausibly falls within the ambit of the First Amendment," and added:

> That plaintiffs wish to scrape data from websites rather than manually record information does not change the analysis. Scraping is merely a technological advance that makes information collection easier; it is not meaningfully different from using a tape recorder instead of taking written notes, or using the panorama function on a smartphone instead of taking a series of photos from different positions. And . . . the information plaintiffs seek is located in a public forum. Hence, plaintiffs' attempts to record the contents of public websites for research purposes are arguably affected with a First Amendment interest.[281]

In the end, the court did not decide whether this First Amendment interest superseded the governmental interests in the CFAA, as the court determined that scraping "fall[s] outside of the CFAA's reach" altogether.[282] The court expressly adopted the "narrow" view of the CFAA discussed in Section III.B. above, and found that under this view, "[s]craping or otherwise recording data from a site that is accessible to the public is merely a particular use of infor-

[275] *See Oracle*, 191 F. Supp. 3d at 1140 (noting that Rimini "continue[d] to download files" despite the blocks and warnings).

[276] *See Power Ventures*, 844 F.3d at 1069; Craigslist Inc. v. 3Taps Inc., 964 F. Supp. 2d 1178, 1185 (N.D. Cal. 2013).

[277] No. 16-cv-1368 (JDB), 2018 WL 1568881(D.D.C. Mar. 30, 2018).

[278] *Id*. at *2. In the interest of full disclosure, I know some of the plaintiffs through our mutual academic circles.

[279] *See Id*.

[280] *Id*. at *7.

[281] *Id*.

[282] *Id*. at *16.

mation that plaintiffs are entitled to see."[283] The court's holding in *Sandvig* rendered this particular technique of research outside the ambit of ongoing constitutional challenge to the CFAA, but in a way that limited application of the CFAA to scraping altogether.

Because *Sandvig* presented a pre-enforcement challenge to the statute the court was not required to consider a situation where the subject of the research might have caught wind of the scraping and issued a letter expressly revoking access. But based on the analysis applied by the court, it would seem that such claims would face greater scrutiny than the purpose-blind approach taken by the courts in the cases in Section III.C. The *Sandvig* court notes, citing a recent Supreme Court case, that the public-facing Internet is "too heavily suffused with First Amendment activity, and what might otherwise be deemed private spaces are too blurred with expressive spaces, to sustain a direct parallel to the physical world,"[284] and because the public should have a general right to access publicly-facing websites, only code-based controls should be the basis of CFAA liability.[285] This approach would perfectly align the CFAA with the technical realities of web scraping described in Section II, though it may not fully answer the question of how to deal with technically-imposed blocks that are motivated by a speech-suppressing purpose, such as an IP block placed to prevent a critic from scraping data from a public website. It again raises the question – first raised by the First Circuit fifteen years ago – whether an even greater public policy limitation should inform CFAA claims based on generally-public websites.[286]

CONCLUSION

At this point it would be absurd to suggest that web scraping could, or should, be generally prohibited. Indeed, many forms of web scraping provide important benefits to consumers and the public.[287] But the legal status of web scraping has gone through twenty years of uncertainty — not a single, incoherent mess as some scholars have suggested, but a status that has ebbed and flowed at different points. After its broad application for about a decade, courts narrowed application to the CFAA, which then gave way to broadening by means of judicial-adoption of the "revocation" theory,[288] and now recent deci-

[283] *Id.*

[284] *Id.* at *5 (citing Packingham v. North Carolina, 137 S. Ct. 1730 (2017)).

[285] *Id.* ("[S]imply placing contractual conditions on accounts that anyone can create, as social media and many other sites do, does not remove a website from the First Amendment protections of the public Internet. If it did, then *Packingham*—which examined a law that limited access to websites that require user accounts for full functionality—would have come out the other way.").

[286] *See supra* notes 171-173 and accompanying text; EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58 (1st Cir. 2003).

[287] *See, e.g.*, GOOGLE, https://www.google.com/ (last visited Mar. 8, 2018).

[288] *See supra* Parts III.A. - III.C.

sions have narrowed the CFAA once more —recognizing that both the public interest in public web scraping and the technical similarities between web scraping and web browsing should limit application of the CFAA to web scraping.[289]

As web scraping litigation enters its third decade, there are a few key issues that courts will have to resolve in order to bring further clarity regarding the CFAA to scrapers and websites alike.

First, and perhaps most obviously, courts should determine how to resolve the tension presented in the "revocation" cases discussed in Section III.C and those finding a more general right to access and scrape in Section III.D. This may at heart mean that courts will need to develop a coherent method of dealing with *conflicting authorizations* to a computer under the CFAA, across different mechanisms. For example, many of these cases present situations where a scraper's access to a website is authorized under code-based mechanisms (e.g., the website's server is online and configured to receive and process HTTP requests from the scraper) but unauthorized through another mechanism (e.g., the website sent the scraper a letter that says they are no longer welcome). It is not clear whether courts have fully confronted conflicting authorization under the CFAA, and established a means of mitigating such authorizations.[290] The closest they have come in web scraping cases is in situations where the scraper uses a third party's account to access the site with permission of the account holder, but not of the website. Those cases tend to favor the website's authorization over the account holder's authorization, but without much consideration of the question.[291]

Second, there is an authorization mechanism that is widely used on the technical side yet conspicuously absent from the legal discussion, which should be brought into the analysis. Courts in scraping cases have yet to meaningfully consider what to do with the existing quasi-technical tool that websites and scrapers have used to broker a relationship for the past two decades: the Robots Exclusion Standard, or "robots.txt" standard.[292] This standard provides a vehicle for websites to express whether or not they wish to allow scrapers on their website, where on the website scrapers should be included or excluded, and whether the HTTP queries the scraper generates should be slowed to avoid overwhelming a website.[293] Under this protocol a website operator can place a text file in the top-level directory of its website, entitled "robots.txt," and then set forth its access rules for scrapers in a source-code-like language readable to

[289] *See supra* Part III.D.

[290] For a recent case unpacking questions of the proper party of authorization, see Philips Med. Sys. P.R. Inc. v. GIS Partners Corp., 203 F. Supp. 3d 221, 234–35 (D.P.R. 2016).

[291] *See supra* notes 164, 202–204, 249–251. For scholarly discussion of this question, see Kerr, *Norms*, *supra* note 116, at 1178–80.

[292] *See* GOURLEY & TOTTY, *supra* note 37, at 229–35.

[293] *See id*.

humans and scrapers alike.[294] Most instructional literature for scrapers encourages them to follow this protocol when scraping.[295] Courts in copyright cases have begun analyzing whether adherence to this protocol is in effect a license that allows the scraper to make the copy it inherently generates when scraping.[296] But few courts have considered its application under the CFAA, and none have gone so far as to suggest that it can be used to demonstrate authorized access to a website.[297] Bringing this commonly-employed tool into the analysis of the CFAA may provide web scrapers with guidance that they can more easily understand and effectuate.

And finally, courts may need to more fully consider whether the publicly-accessible nature of most popular websites compel courts to review CFAA claims more carefully where such sites revoke access for a single person — whether the sites revoke access by way of a cease-and-desist letter, a technical block, or any other mechanism. As the *Sandvig* court noted, there is a fluid relationship between private websites and the public web, and many socially beneficial reasons that a person may want to scrape another's website without permission in this quasi-public space.[298] And because most scrapers, if designed appropriately, would be highly similar to the level of access of a human browser,[299] courts should raise an eyebrow at a claim that a scraper should be

[294] *About robots.txt*, WEB ROBOTS PAGES, http://www.robotstxt.org/robotstxt.html [https://perma.cc/7YLT-XDCD] (last visited Mar. 8, 2018).

[295] *See, e.g.*, HEMENWAY & CALISHAIN, *supra* note 37, at 46 ("If you're planning on releasing your scraper or spider into the wild, it's important that you make every possible attempt to support robots.txt.").

[296] *See, e.g.*, Parker v. Yahoo!, Inc., No. 07-2757, 2008 WL 4410095, at *4 (E.D. Pa. Sept. 25, 2008) (suggesting, but not deciding, that the knowing omission of a robots.txt file could be grounds to establish an implied license under copyright law); Field v. Google Inc., 412 F. Supp. 2d 1106, 1117 (D. Nev. 2006) (plaintiff's knowing use of the robots.txt protocol estopped him from arguing that defendant's copying was infringement); *but see* Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537, 563–64 (S.D.N.Y. 2013) (rejecting use of robots.txt to claim an implied license, when the copyright owner was not the operator of the websites in question).

[297] *See* DHI Grp., Inc. v. Kent, No. H-16-1670, 2017 WL 4837730, at *5 (S.D. Tex. Oct. 26, 2017) (using the protocol to inform plaintiff's anticircumvention claim, but not its CFAA claim); QVC, Inc. v. Resultly, LLC, 159 F. Supp. 3d 576, 586 n.3 (E.D. Pa. 2016) (in the same case, rejecting claim that plaintiff's prior statements discussing the robots.txt protocol meant it was judicially estopped from claiming access without authorization); QVC, Inc. v. Resultly, LLC, 99 F. Supp. 3d 525, 539 (E.D. Pa. 2015) (noting that defendant adhered to "crawl delays" articulated in a robots.txt file, but not using the protocol to inform the question of authorized access); Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey, 497 F. Supp. 2d 627, 643 (E.D. Pa. 2007) (extensively analyzing the protocol as it relates to plaintiff's claim under copyright's anticircumvention law, but not as it relates to the plaintiff's CFAA claim); eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1062 (N.D. Cal. 2000) (mentioning the protocol, but not applying it in that case).

[298] Sandvig v. Sessions, No. 16-cv-1368 (JDB), 2018 WL 1568881 at * 5 (Mar. 30, 2018).

[299] *See supra* Part II.

viewed as an invasive criminal trespasser. It may be that the platform's true motivations are actually anticompetitive, speech-suppressing, or otherwise untrustworthy. As the *hiQ Labs* decision put it:

> Website owners could, for example, block access by individuals or groups on the basis of race or gender discrimination. Political campaigns could block selected news media, or supporters of rival candidates, from accessing their websites. Companies could prevent competitors or consumer groups from visiting their websites to learn about their products or analyze pricing. Further . . . [a] broad reading of the CFAA could stifle the dynamic evolution and incremental development of state and local laws addressing the delicate balance between open access to information and privacy—all in the name of a federal statute enacted in 1984 before the advent of the World Wide Web.[300]

∴

There are countless uses of web scraping. Some are good. Some are bad. Some are bad for the website but should be allowed for the good of the public.[301] The past twenty years of CFAA web scraping litigation have slowly worked their way towards a broader appreciation of the nature and potential benefits of scraping. A more detailed look at the actual technical function of scraping shows that scraping should not be thought of as inherently more invasive or dangerous than a person at a web browser. And with a few key areas of doubt cleared away, one can hope that the constant legal uncertainty expressed by coders, lawyers, and scholars alike will give way to greater coherence and clarity.

[300] hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099, 1110–1111 (N.D. Cal. 2017).

[301] *See supra* notes 13–25 and accompanying text.