ACCESS GRANTED: A FIRST AMENDMENT THEORY OF REFORM OF THE CFAA ACCESS PROVISION

*Jacquellena Carrero**

Data scraping—the automated collection of data on the internet—is used in a variety of contexts. On the commercial side, scraping might be used as a means of competition—such as scraping by one company to retrieve information on prices for services provided by a competitor. On the noncommercial side, scraping could be used as a research tool—such as scraping by a news outlet to investigate Amazon's pricing algorithm. Despite the varied applications of data scraping, courts' varying interpretations of the Computer Fraud and Abuse Act (CFAA) can impose both civil and criminal liability for scraping. This Note argues that there are competing First Amendment interests both in favor of and against scraping, depending on the type of scraping conducted. Because the CFAA does not distinguish between various types of scraping and balance these competing interests, a legislative solution is needed to comport with both First Amendment interests of accountability and political self-governance on one end, and privacy on the other end.

INTRODUCTION

In the wake of the news that Russia had meddled in the 2016 presidential election,[1] researcher Jonathan Albright decided to explore the reach of the Russian disinformation campaign on Facebook.[2] In early October, around the time Albright was conducting his research on Russian disinformation-campaign networks, Facebook claimed that politically divisive advertisements purchased by Russian operatives had reached ten million

* J.D. Candidate 2020, Columbia Law School. The author would like to thank Professor David Pozen, Ramya Krishnan and Carrie DeCell at the Knight First Amendment Institute, and the staff of the *Columbia Law Review* for their invaluable guidance and thoughtful contributions to this Note. Special thanks to her family and her partner, Johnny Araujo, for their endless support and encouragement.

1. See Office of the Dir. of Nat'l Intelligence, ICA 2017-01D, Assessing Russian Activities and Intentions in Recent US Elections, at ii (2017), https://www.dni.gov/files/documents/ICA_2017_01.pdf [https://perma.cc/L35B-TFTM] ("We assess Russian President Vladimir Putin ordered an influence campaign in 2016 aimed at the US presidential election.").

2. See Craig Timberg, Russian Propaganda May Have Been Shared Hundreds of Millions of Times, New Research Says, Wash. Post (Oct. 5, 2017), https://www.washingtonpost.com/news/the-switch/wp/2017/10/05/russian-propaganda-may-have-been-shared-hundreds-of-millions-of-times-new-research-says/ (on file with the *Columbia Law Review*).

users in the months before and after the 2016 election.[3] But Facebook's ten million figure accounted only for paid advertisements; it failed to include free accounts created by the Russians that influenced Facebook's massive user base.[4] Albright theorized that the true reach of Russian meddling encompassed "all the activity of the Russian controlled accounts—each post, each 'like,' each comment"—in addition to the paid advertisements.[5] His hypothesis proved true; the actual extent of Russia's influence may have been "well into the billions of 'shares' on Facebook."[6] Following journalists' reports about how many Americans were exposed to Russian propaganda, Facebook revised its number to 150 million people affected.[7] Albright's groundbreaking research would not have been possible without the internet-research technique known as "scraping"—that is, the automated "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."[8] Albright used the Facebook-owned analytics tool CrowdTangle to automatically download the five hundred most recent posts for each of the Russian campaign accounts and analyze their reach.[9]

Albright's exposé of the true extent of the Russian misinformation campaign on Facebook is consistent with the First Amendment values of democratic self-governance, democratic legitimation, truth, and autonomy.[10] For example, legal scholar Alexander Meiklejohn argued that self-

---

3. David Ingram, Facebook Says 10 Million U.S. Users Saw Russia-Linked Ads, Reuters (Oct. 2, 2017), https://www.reuters.com/article/us-facebook-advertising/facebook-says-10-million-u-s-users-saw-russia-linked-ads-idUSKCN1C71YM [https://perma.cc/BMT2-XDYE].

4. See Timberg, supra note 2.

5. Id.

6. Id.; see also Nicholas Confessore & Daisuke Wakabayashi, How Russia Harvested American Rage to Reshape U.S. Politics, N.Y. Times (Oct. 9, 2017), https://www.nytimes.com/2017/10/09/technology/russia-election-facebook-ads-rage.html (on file with the *Columbia Law Review*) (citing Albright's research to describe how Russia's social media campaign used both paid advertisements and regular, nonpaid posting features of Facebook to influence voters).

7. Spencer Ackerman, Facebook Now Says Russian Disinfo Reached 150 Million Americans, Daily Beast (Nov. 11, 2017), https://www.thedailybeast.com/facebook-now-says-russian-disinfo-reached-150-million-americans [https://perma.cc/4YLX-FMTJ].

8. Andrew Sellars, Twenty Years of Web Scraping and the Computer Fraud and Abuse Act, 24 B.U. J. Sci. & Tech. L. 372, 373 (2018). An application programming interface is a set of "requirements that govern how one application can talk to another" and enable the movement of information from one program to another. Brian Proffitt, What APIs Are and Why They're Important, ReadWrite (Sept. 19, 2013), https://readwrite.com/2013/09/19/api-defined/ [https://perma.cc/TEP4-H3S6].

9. Timberg, supra note 2. For the results of Albright's research conducted via scraping mechanisms, see Jonathan Albright (d1gi), Itemized Posts and Historical Engagement—6 Now-Closed FB Pages, Tableau Public, https://public.tableau.com/profile/d1gi#!/vizhome/FB4/TotalReachbyPage (on file with the *Columbia Law Review*) (last updated Oct. 5, 2017).

10. See Jeremy K. Kessler & David E. Pozen, The Search for an Egalitarian First Amendment, 118 Colum. L. Rev. 1953, 1978–79 (2018) ("[J]udges and scholars have

government justifies free speech because speech and the resulting exchange of information produce informed voters.[11] Thus, asserting access to information that affects democracy and elections is "paramount to self-governance."[12] Similarly, First Amendment scholar Seana Shiffrin posited an autonomy-based theory of free speech: Autonomous thinkers should have access to information in order to be able to think freely and reflect.[13] This need for access to information becomes only more pronounced in today's ever-expanding digital society as big data, technology companies, and social media platforms are transforming public discourse and influencing democracy in ways that are often obscured from the public eye.[14]

Albright's research demonstrates a key tension between journalists and the technology companies they seek to investigate: While there is a First Amendment interest in scraping, these techniques can also subject researchers and journalists to legal liability. Following Albright's research, Facebook fixed the glitch in the CrowdTangle tool that had allowed him to access the data for his research.[15] A spokesperson stated that the scrap-

---

produced a vast body of writing that seeks to justify, critique, and shape First Amendment doctrine in light of foundational principles and aspirations—above all, the pursuit of truth, the promotion of individual autonomy, and the facilitation of democratic self-government.").

11. See Alexander Meiklejohn, Free Speech and Its Relation to Self-Government 3–8, 22–27 (1948); Alexander Meiklejohn, The First Amendment Is an Absolute, 1961 Sup. Ct. Rev. 245, 255–57, 263 [hereinafter Meiklejohn, First Amendment Absolute] ("'[T]he people need free speech' because they have decided, in adopting, maintaining and interpreting their Constitution, to govern themselves rather than to be governed by others." (quoting Harry Kalven, Jr., Metaphysics of the Law of Obscenity, 1960 Sup. Ct. Rev. 1, 16)). Subsequent legal scholarship has drawn on Meiklejohn's theory of information as a key component of a functioning democracy. See, e.g., Thomas L. Emerson, Legal Foundations of the Right to Know, 1976 Wash. U. L.Q. 1, 2 ("[T]he right to know . . . is a significant method for seeking the truth, or at least for seeking the better answer.").

12. D. Victoria Baranetsky, Data Journalism and the Law, Tow Ctr. for Dig. Journalism (Sept. 19, 2018), https://www.cjr.org/tow_center_reports/data-journalism-and-the-law.php/ [https://perma.cc/23G5-MXZA] (describing the ways in which journalists' access to critical data is being restricted in an environment oversaturated with information).

13. See Seana Valentine Shiffrin, A Thinker-Based Approach to Freedom of Speech, 27 Const. Comment. 283, 289–92 (2011).

14. See, e.g., Bernard Marr, Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance 43–44 (2015) (ebook) ("There is little doubt that [b]ig [d]ata is changing the world. It is already completely transforming the way we live, find love, cure cancer, conduct science, improve performance, run cities and countries and operate business."); Alex Abdo, Facebook Is Shaping Public Discourse. We Need to Understand How, Guardian (Sept. 15, 2018), https://www.theguardian.com/commentisfree/2018/sep/15/facebook-twitter-social-media-public-discourse [https://perma.cc/B2JH-HYF3] ("Facebook's alluring user interface obscures an array of ever-changing algorithms that determine which information you see and the order in which you see it. The algorithms are opaque—even to Facebook—because they rely on a form of computation called 'machine learning', in which the algorithms train themselves . . . .").

15. Natasha Bertrand, Facebook Scrubbed Potentially Damning Russia Data Before Researchers Could Analyze It Further, Bus. Insider (Oct. 12, 2017), https://www.

ing mechanisms Albright used were "an unintended way to access information about deleted content."[16] Although Facebook did not take legal action against Albright for his scraping activity, they potentially could have under the Computer Fraud and Abuse Act (CFAA).[17] The CFAA is a cybersecurity statute aimed at penalizing misuse of private information and damage to computers.[18] Many websites contain provisions in their terms of service that effectively prohibit web scraping,[19] and some courts have interpreted the CFAA to penalize such violations of a website's terms of service.[20]

The CFAA may be overbroad and overinclusive, but it does capture problematic uses of data; while researchers and journalists may be harnessing data in a way that arguably promotes accountability and democratic participation, others have harnessed data for different motivations not affected with a First Amendment interest. For example, Facebook recently faced scrutiny over the data firm Cambridge Analytica's harvesting of users' personal information to target users with personalized political advertisements ahead of the 2016 presidential election.[21] Therefore, the

businessinsider.com/facebook-russia-data-fake-accounts-2017-10 [https://perma.cc/34EJ-MCTK].

16. Id. (internal quotation marks omitted).

17. 18 U.S.C. § 1030(a)(2)(C) (2012) (prohibiting access to "information from any protected computer").

18. Congress's main goal in enacting the CFAA as the first federal cybercrime law was to combat "so-called 'hackers' who have been able to access (trespass into) both private and public computer systems." H.R. Rep. No. 98-894, at 10 (1984); see also Mark A. Lemley, Place and Cyberspace, 91 Calif. L. Rev. 521, 528 (2003) (noting that the CFAA "was designed to punish malicious hackers"). The CFAA was inspired by the 1983 film *WarGames* after President Ronald Reagan screened the film at Camp David and instructed his advisors to look into securing government computers from hacking. See Gabe Rottman, Knight Institute's Facebook 'Safe Harbor' Proposal Showcases Need for Comprehensive CFAA Reform, Reporters Comm. for Freedom of the Press (Aug. 6, 2018), https://www.rcfp.org/knight-institutes-facebook-safe-harbor-proposal-showcases-need-compr/ [https://perma.cc/E83Y-B389] ("Taken by the film, . . . [President Reagan] interrupted a meeting with his joint chiefs to ask if the scenario was at all realistic. His advisors looked into it . . . and recommended immediate action . . . ."); see also H.R. Rep. No. 98-894, at 10 (noting that *WarGames* showed "a realistic representation of the automatic dialing and access capabilities of the personal computer" (quoting Counterfeit Access Device and Computer Fraud and Abuse Act: Hearing on H.R. 3181, H.R. 3570, and H.R. 5112 Before the Subcomm. on Crime of the H. Comm. on the Judiciary, 98th Cong. 185 (1984) (statement of Peter Waal, Vice President, Marketing, GTE-Telenet))). The law has since been expanded to apply to any computer connected to the internet. See Rottman, supra. Courts have interpreted the CFAA broadly to cover an array of activity beyond strictly hacking. See infra section I.B.

19. See e.g., User Agreement, LinkedIn, https://www.linkedin.com/legal/user-agreement [https://perma.cc/HP6A-GP4W] (last updated May 8, 2018) ("You agree that you will not . . . [d]evelop, support or use software, devices, scripts, robots, or any other means or processes . . . to scrape the Services or otherwise copy profiles and other data . . . .").

20. See infra section I.B (discussing the circuit split on "exceeds authorized access").

21. Carole Cadwalladr & Emma Graham-Harrison, Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach, Guardian (Mar. 17, 2018), https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-

CFAA can both serve as a recourse for improper uses of data—as with the Cambridge Analytica scandal—and present a substantial obstacle to researchers and journalists seeking to perform research that serves the public interest, as was Albright. Thus, the CFAA crudely lumps together different forms of scraping that have different motivations and implications for social values.

This Note assesses First Amendment interests with regard to scraping and argues that there are competing First Amendment interests both in favor of and against scraping. While scraping that serves the public interest merits First Amendment protection, commercially oriented scraping can threaten First Amendment values of intellectual privacy and should not receive the same protections. Part I provides background on data-scraping techniques, taxonomizes the various applications of data scraping, and outlines how courts' varying interpretations of the CFAA can impose liability for data scraping. Next, Part II argues that there are competing First Amendment and privacy interests in data-scraping activity as implicated by the CFAA. Finally, Part III proposes that because the CFAA does not appropriately balance these competing interests, a legislative solution is needed so that the law comports with both First Amendment interests of accountability and political self-governance on one end, and privacy on the other end.

I. DATA SCRAPING, THE CFAA, AND THE FIRST AMENDMENT

This Part introduces, in more detail, the different contexts in which data scraping arises and explains how these practices can result in liability under the CFAA. Section I.A discusses the recent phenomenon of data-scraping mechanisms and provides examples of scraping in academia, journalism, and other fields to give an overview of these techniques and their use. Section I.B turns to recent jurisprudence on such instances of scraping and describes how judicial interpretations of the CFAA access provision[22] can result in liability for engaging in data scraping.

---

influence-us-election [https://perma.cc/HMZ8-WCMZ] (describing how the data analytics firm linked to former Trump advisor Steve Bannon "compiled user data to target American voters").

22. Each of the CFAA sections imposing liability contains an "access provision" stipulating that users who access information "without authorization" or who "exceed[] authorized access" are subject to liability. See 18 U.S.C. § 1030(a)(2)–(5). For further explanation of the CFAA access provisions and courts' interpretations, see infra notes 126–131.

A. Data Scraping: A Taxonomy

In an era in which society creates floods of data every day[23] via computers, GPS devices, and cell phones, data sets are increasingly perceived as a "new class of economic asset, like currency or gold."[24] Some have described "big data" as "the data of the Web—online searches, posts and messages."[25] Others have defined the term as "massive quantities of information produced by and about people, things, and their interactions."[26] This Note uses big data as a "shorthand for advancing trends in technology that open the door to a new approach to understanding the world and making decisions."[27]

The increasing prominence of big data has led to a desire to harvest it for a variety of purposes, which has subsequently given rise to methods of harnessing data, like web scraping.[28] To lay the foundation for the proposed CFAA reform solution presented in Part III, section I.A.1 describes web scraping and outlines a few techniques of information retrieval commonly used by scrapers. The following section then attempts to taxonomize the various applications of web scraping.[29] Section I.A.2.a discusses web scraping as it has been utilized by commercial actors. Section I.A.2.b discusses web scraping as it has been utilized by noncommercial actors whose purpose is journalistic, scholarly, or scientific research.

1. *Web Scraping: Definition and Techniques.* — Web scraping, a method of harnessing data, mimics human browsing to achieve "programmatic browsing of web pages and targeted data collection" without human interaction and with greater precision than its predecessor methods.[30] Most methods of scraping are conducted via a "computer script that will

---

23. See Bernard Marr, How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read, Forbes (May 21, 2018), https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read [https://perma.cc/95G5-CB4C] (noting that "90 percent of the data in the world" was created from 2016 to 2018).

24. Steve Lohr, The Age of Big Data, N.Y. Times (Feb. 11, 2012), http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html (on file with the *Columbia Law Review*).

25. Id.

26. danah boyd & Kate Crawford, Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon, 15 Info. Comm. & Soc'y 662, 663 (2012).

27. Lohr, supra note 24.

28. See id. (describing the potential applications of big data for retailers, businesses, online dating services, police departments, and researchers).

29. The classifications of noncommercial, commercial, and other uses of web scraping are borrowed from the Freedom of Information Act (FOIA) context. See infra notes 61–64 and accompanying text (describing the FOIA classifications and how they are utilized in this Note).

30. See Nicholas A. Wolfe, Hacking the Anti-Hacking Statute: Using the Computer Fraud and Abuse Act to Secure Public Data Exclusivity, 13 Nw. J. Tech. & Intell. Prop. 301, 305 (2015).

send tailored queries to websites to retrieve specific pieces of content."[31] These requests are usually automated in order to cover a large span of data from a specific website or array of websites.[32] Web scraping typically collects data from "screen outputs or . . . the HyperText Markup Language ('HTML') code that most websites display."[33] The tools of scraping come in different formats, but some common methods are browser extension,[34] installable software,[35] and cloud-based technology.[36] The advantage of web scraping is that it speeds up the tedious job of manually copying and pasting data into a spreadsheet, making large-scale data collection possible by automating the process.[37]

Because web scraping practices encompass a broad range of activity, courts have not come to a consensus on common terminology for web scraping or what activity qualifies as "scraping."[38] In 1996, the Southern District of New York was the first court to define "web scraper," defining the term as "software capable of automatically contacting various Web sites and extracting relevant information."[39] However, a more precise and robust definition can be found in the First Circuit's opinion in *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.[40]

---

31. Sellars, supra note 8, at 373.

32. Id. at 373–74.

33. Jeffrey Kenneth Hirschey, Symbiotic Relationships: Pragmatic Acceptance of Data Scraping, 29 Berkeley Tech. L.J. 897, 897–98 (2014).

34. See Prowebscraper, Types of Web Scraping Tools, Medium (Mar. 7, 2018), https://medium.com/prowebscraper/types-of-web-scraping-tools-940f824622fb [https://perma.cc/K5AM-D9DY] (describing how extension scraping tools can scrape data from one website page at a time and provide the data in a downloadable format).

35. See id. (describing installable software as the best option for scraping a medium amount of data from multiple web pages at a time).

36. See id. (describing cloud-based tools that can scrape large amounts of data).

37. What Is Web Scraping?, WebHarvy, https://www.webharvy.com/articles/what-is-web-scraping.html [https://perma.cc/8XNS-WDC9] (last visited Sept. 13, 2019).

38. 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."). For a thorough summary of the terminology used to refer to web scraping in court opinions, see Sellars, supra note 8, at 381–82.

39. Sellars, supra note 8, at 383 (internal quotation marks omitted) (quoting Shea ex rel. Am. Reporter v. Reno, 930 F. Supp. 916, 929 (S.D.N.Y. 1996)).

40. 318 F.3d 58, 60 (1st Cir. 2004).

However, as legal scholar Andrew Sellars has noted, this definition is limited in two ways.[41] First, it does not include the common use of web scraping to "follow links around to other websites hosted on other computers."[42] Second, the First Circuit's definition does not incorporate dynamic websites that are "generated as they are requested."[43] The content on most modern websites is rarely preloaded, statically generated information. Instead, websites utilize the user's account information, geographic location, and type of device used to create content.[44] In fact, one purpose of web scraping is to determine how "inputs can change the outputs."[45]

Existing scholarship has comprehensively documented the rise of algorithms that crunch big data "culled from individuals' online activity."[46] For example, Apple's personal-assistant application Siri learns from past answers to be able to respond to an "expanding universe of questions."[47] Importantly, such algorithms can also develop to provide discriminatory results.[48] Researchers examine discriminatory algorithms by conducting a scraping audit in which a "researcher uses a bot to impersonate users of various backgrounds."[49] The bot utilized by the scraper "issues repeated queries" to monitor and record the algorithm's response.[50] For example, a group of Harvard Business School researchers used the scraper-audit technique to examine instances of racial discrimination on the online short-term rental marketplace service Airbnb.[51] The researchers constructed user profiles with black-sounding names and white-sounding names

41. See Sellars, supra note 8, at 384.

42. Id.; see also Kevin Hemenway & Tara Calishain, Spidering Hacks: 200 Industrial-Strength Tips & Tools 18–19 (2004).

43. Sellars, supra note 8, at 384.

44. Id.

45. Id.

46. Komal S. Patel, Note, Testing the Limits of the First Amendment: How Online Civil Rights Testing Is Protected Speech Activity, 118 Colum. L. Rev. 1473, 1474 (2018); see also Lohr, supra note 24 ("The wealth of new data, in turn, accelerates advances in computing — a virtuous circle of Big Data. Machine-learning algorithms, for example, learn on data, and the more data, the more the machines learn.").

47. Lohr, supra note 24.

48. One recent example of this phenomenon occurred in 2016 when Amazon faced scrutiny because its Prime membership same-day delivery service excluded predominantly black ZIP codes. See David Ingold & Spencer Soper, Amazon Doesn't Consider the Race of Its Customers. Should It?, Bloomberg (Apr. 21, 2016), https://www.bloomberg.com/graphics/2016-amazon-same-day/ [https://perma.cc/S5NT-N2XX]. For a brief introduction to discriminatory algorithm practices and a thorough discussion of the promise and pitfalls surrounding algorithms in the lending context, see Matthew Adam Bruckner, The Promise and Perils of Algorithmic Lenders' Use of Big Data, 93 Chi.-Kent L. Rev. 3, 17–29 (2018).

49. Patel, supra note 46, at 1474.

50. Id.

51. See Ray Fisman & Michael Luca, Fixing Discrimination in Online Marketplaces, Harv. Bus. Rev. (Dec. 2016), http://hbr.org/2016/12/fixing-discrimination-in-online-marketplaces [https://perma.cc/BPC3-RLA8].

and sent rental requests to over six thousand Airbnb hosts to test how responses differed based on race.[52] As these examples demonstrate, scraping is much more than a computer simply downloading static web results like a human would, except at a much faster speed.[53]

As with the definition of scraping, courts' perceptions of and attitudes toward scrapers has varied. Some judges have portrayed scrapers as nefarious actors, such as a vandal "taking a hammer to a piece of machinery"[54] or a person walking into a bank with both a safe deposit key and a shotgun.[55] More innocently, courts have also likened scrapers to a roving photographer for the internet,[56] a person "using a tape recorder instead of taking written notes,"[57] and a person using the panoramic function on a cell phone instead of a taking a series of photos.[58] Similarly, the uses of web scraping are as diverse as the reasons for harvesting data.[59] Some examples include companies monitoring information posted on social media to stay abreast of issues requiring customer support, e-commerce businesses monitoring competitors' pricing and inventory, and manufacturers "tracking the performance ranking of products in the search results of retailer websites."[60]

2. *Scraping Taxonomy.* — This section attempts to taxonomize instances of scraping to provide a comprehensive overview of scraping techniques. This Note makes a distinction between commercial and noncommercial uses of web scraping. Other transparency frameworks employ a similar commercial–noncommercial distinction to categorize information requests. For example, requests for government information under the Freedom of Information Act (FOIA) are classified for fee determinations based on the identity of the requesters and their intended use of the

---

52. Id. For further discussion of scraping techniques to test online discrimination, see infra section I.A.2.b.

53. Sellars, supra note 8, at 384.

54. Ticketmaster Corp. v. Tickets.com, Inc., No. 99CV7654, 2000 WL 1887522, at *4, (C.D. Cal. Aug. 10, 2000).

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

56. See Healthcare Advocates, Inc. v. Harding, Earley, Follmer & Frailey, 497 F. Supp. 2d 627, 631 (E.D. Pa. 2007) (describing scraping as "an automated program that scours the Internet and takes pictures of every web page that it is instructed to visit").

57. Sandvig v. Sessions, 315 F. Supp. 3d 1, 16 (D.D.C. 2018).

58. Id.

59. See Jamie Williams, 'Scraping' Is Just Automated Access, and Everyone Does It, Elec. Frontier Found. (Apr. 17, 2018), https://www.eff.org/deeplinks/2018/04/scraping-just-automated-access-and-everyone-does-it [https://perma.cc/2WY6-HLCM] ("Companies use automated web browsing products to gather web data for a wide variety of uses.").

60. Id.; see also Doug Laney, Gartner Predicts Three Big Data Trends for Business Intelligence, Forbes (Feb. 12, 2015), https://www.forbes.com/sites/gartnergroup/2015/02/12/gartner-predicts-three-big-data-trends-for-business-intelligence/ [https://perma.cc/X2LV-MJDZ] ("Your company's biggest database isn't your transaction . . . or other internal database. Rather it's the Web itself and the world of exogenous data now available from syndicated and open data sources.").

information.[61] Since this Note argues that web scraping serves First Amendment values when the information sought to be obtained serves the public interest,[62] the commercial–noncommercial classification based on the identity of the requester and the purpose of the request is one method of prioritizing journalists and researchers.[63] The definitions of commercial and noncommercial use are borrowed from the FOIA context, since FOIA is a similar transparency framework and these definitions have had the benefit of being tested and refined by the case law.[64]

a. *Commercial Web Scraping.* — The first category in this Note's scraping taxonomy is scraping for commercial uses. With respect to FOIA requests,

---

61. See 5 U.S.C. § 552(a)(4)(A)(ii) (2012); FOIA Guide: Fees and Fee Waivers, Dep't of Justice, https://www.justice.gov/oip/foia-guide-2004-edition-fees-and-fee-waivers [https://perma.cc/AL9C-2V9R] [hereinafter FOIA Guide] (last updated July 23, 2014) (describing the three fee levels imposed on various types of FOIA requests). The three FOIA fee classifications are: (1) requested for commercial use; (2) requested not for commercial use and "the request is made by an educational or noncommercial scientific institution, whose purpose is scholarly or scientific research; or a representative of the news media;" and (3) any other request not described. See id. It is important to note that while FOIA categorizes requests for fee purposes, the classification of the request has no bearing on how it is processed. See 5 U.S.C. § 552(a)(4)(A)(ii). That is—other than the fees and processing time—FOIA requests are responded to in an identical manner regardless of whether the requester is a commercial or noncommercial user. See id.; see also Margaret B. Kwoka, First-Person FOIA, 127 Yale L.J. 2204, 2256 (2018) ("FOIA was enacted expressly to disavow any restriction based on identity or purpose, precisely because this limitation operated so poorly that agencies used it as an excuse to deny access arbitrarily.").

62. See infra section II.A.

63. Legal scholars have argued that FOIA should adopt a context-specific approach and consider the identity of the information requester to more efficiently serve its goals of democratic oversight and accountability. See, e.g., Erin C. Carroll, Protecting the Watchdog: Using the Freedom of Information Act to Preference the Press, 2016 Utah L. Rev. 193, 195 ("Providing faster and better access to information about government activity would feed investigative journalism and give the press the heft it needs to better serve as a check against government . . . ."); Kwoka, supra note 61, at 2211. Indeed, while in theory FOIA is intended to promote accountability, in practice FOIA "systematically advantages certain private concerns as well as certain blocs within government," such as commercial requesters, contractors, and lawyers. David E. Pozen, Freedom of Information Beyond the Freedom of Information Act, 165 U. Pa. L. Rev. 1097, 1112–23 (2017) [hereinafter Pozen, Beyond FOIA] (describing the structural shortcomings of FOIA that result in commercially interested parties receiving special benefits). FOIA's fee structure is instructive since limiting fees based on the identity of the requester—even while not differentiating processing of requests—is one method of preferencing journalists and researchers. See id. at 1140 ("FOIA . . . preferences the press . . . by limiting fees to 'reasonable standard charges for document duplication when the requester is from the news media (or 'an educational or non-commercial scientific institution'), and by providing for expedited processing of requests 'made by a person primarily engaged in disseminating information.'" (footnote omitted) (quoting 5 U.S.C. § 552(a)(4)(A)(ii)(II); then quoting 5 U.S.C. § 552(a)(6)(E)(i), v(II))). However, this Note will suggest more aggressive preferential treatment for journalists and researchers in the scraping context. See infra Part III (suggesting legislative reform of the CFAA to create a statutory safe harbor provision).

64. See infra notes 90–91 and accompanying text.

a "commercial use" is defined as "a use or purpose that furthers the commercial, trade or profit interests of the requester or the person on whose behalf the request is being made."[65] The classification of a commercial user of scraping techniques for the purposes of this taxonomy turns on the intended use of the information, rather than the identity of the requester.[66]

Scraping tactics and subsequent disputes arise most frequently in the commercial use context. Sellars collected all of the cases involving web scraping and identified sixty-one decisions involving methods of scraping.[67] In his survey, he noted that a vast number of these claims are brought by direct commercial competitors or "companies in closely adjacent markets to each other."[68] The type of scraping at issue in *EF Cultural Travel BV v. Zefer Corp.* is one example of scraping as a means of commercial competition.[69] In that case, student-travel services company Explorica used a robot scraping program designed by computer services company Zefer to crawl EF Cultural Traveler's website to glean information on its competitor's tour prices.[70] The scraper program accessed information from the HTML source code that generated the graphical interface—which was available to anyone who viewed the site—and downloaded that information to the user's computer.[71] The scraping tool did not copy all the information available on the scraped pages, but only copied the price for each tour through each possible gateway city.[72] After it received the information on EF Cultural Traveler's price points, Explorica "set its own prices for the public" and "undercut[] EF's prices an average of five percent."[73] *hiQ Labs*

65. OMB Fee Guidelines, 52 Fed. Reg. 10,012, 10,017–18 (Mar. 27, 1987); see also Avondale Indus. v. NLRB, No. Civ.A. 96–1227, 1998 WL 34064938, at *4–5 (E.D. La. Mar. 20, 1998) (embracing OMB's definition of "commercial use").

66. See OMB Fee Guidelines, 52 Fed. Reg. at 10,018. To see how this classification based on intended use has played out in the FOIA context, see Comer v. IRS, No. 97-CV-76329, 1999 WL 1022210, at *4 (E.D. Mich. Sept. 30, 1999) (noting that an agency can consider a requestor's motivations for the "commercial user" determination for FOIA fee purposes); Hosp. & Physician Publ'g v. U.S. Dep't of Def., No. 98-CV-4117, 1999 WL 33582100, at *4–5 (S.D. Ill. June 22, 1999) (suggesting that the key inquiry in commercial use designation is the current intended use and that a requester's past commercial use of records has no bearing on the determination); S.A. Ludsin & Co. v. SBA, No. 96 CV 5972, 1998 WL 355394, at *2 (E.D.N.Y. Apr. 2, 1998) (finding a requester who sought documents to enhance the prospect of securing a government contract to be a commercial requester); *Avondale*, 1998 WL 34064938, at *4–5 (finding a company's intent to use requested documents to contest union election results to be commercial use).

67. See Sellars, supra note 8, at 378 & n.43 (collecting cases and describing the goal of his scholarship as "break[ing] down the twenty years of web scraping litigation").

68. Id. at 390.

69. 318 F.3d 58 (1st Cir. 2003).

70. Id. at 60.

71. Id.

72. Id.

73. Id. The First Circuit affirmed the lower court's ruling that Zefer's scraping violated the CFAA. For further discussion of the First Circuit's approach to CFAA interpretation as it involves scraping, see infra notes 129–131 and accompanying text.

v. LinkedIn demonstrated another commercial use of scraping.[74] In that case, plaintiff hiQ Labs routinely used "bots" to scrape publicly available profiles on LinkedIn to create alerts to employers about their employees' online activity.[75] hiQ's business model depended entirely on access to LinkedIn's public data.[76]

Commercial web scraping can also encompass mixed-motive scraping—scraping that is not purely commercial but has a commercial component. One recent example is the Cambridge Analytica scandal which fused commercial, noncommercial, and political motivations for data collection. In June 2014, Cambridge Analytica, SCL Group Limited (SCL), and Global Science Research Limited (GSR) acquired the personal information of 71.6 million Facebook users, including names, phone numbers, mailing and email addresses, political and religious affiliations, and interests.[77] Facebook granted GSR access to collect data through an app called "thisisyourdigitallife."[78] Through the app, users were prompted to provide their Facebook login credentials to take a quiz.[79] However, the design of the application allowed the app's creator, Aleksander Kogan, to obtain users' personal information as well as data from users' friends.[80] Cambridge Analytica and GSR then used this data "to build a powerful software program to predict and influence choices at the ballot box."[81] The information was "harvested on an unprecedented scale," with "more than 50 million individuals" affected."[82]

Cambridge Analytica's scraping falls within the commercial classification since users' data were sold. Although it could be argued that the use was in the public interest because the information was used for voting influence, the scraping's primary purpose was to secretly influence a democratic election's results. In fact, this was the first known incident of politically motivated data collection that used misinformation to trick

---

74. hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099 (N.D. Cal. 2017), aff'd and remanded, No. 17-16783, 2019 WL 4251889 (9th Cir. Sept. 9, 2019).

75. Id. at 1104.

76. Id. For further discussion of the outcome of *hiQ Labs v. LinkedIn*, see infra notes 279–283 and accompanying text.

77. Fields PLLC, Class Action Lawsuit Filed Against Facebook and Cambridge Analytica for Stealing and Improperly Using More Than 71 Million Users' Data, Cision PR Newswire (Apr. 10, 2018), https://www.prnewswire.com/news-releases/class-action-lawsuit-filed-against-facebook-and-cambridge-analytica-for-stealing-and-improperly-using-more-than-71-million-users-data-300627281.html [https://perma.cc/T7TK-49TH].

78. Id.

79. Id.

80. See Eric Auchard & David Ingram, Cambridge Analytica CEO Claims Influence on U.S. Election, Facebook Questioned, Reuters (Mar. 20, 2018), https://www.reuters.com/article/us-facebook-cambridge-analytica/cambridge-analytica-ceo-claims-influence-on-u-s-election-facebook-questioned-idUSKBN1GW1SG [https://perma.cc/L4TM-7MP7].

81. Cadwalladr & Graham-Harrison, supra note 21.

82. Id.

people into disclosure.[83] For this reason, the Cambridge Analytica scandal is widely viewed as more nefarious than other scraping instances.[84] Since news of the scandal broke, both Facebook and Cambridge Analytica have been the target of multiple class action lawsuits.[85]

It is beyond the scope of this Note to catalogue all the instances of commercial web scraping, but the cases and the existing scholarship discussed herein demonstrate that web scraping is a valuable tool for countless commercial purposes.

b. *Noncommercial Web Scraping for Research and Journalistic Purposes.* — Scraping is also often used for noncommercial purposes by educational and scientific research institutions, as well as journalists investigating a story. To qualify as an "educational institution," FOIA requires that information requesters must be "schools, as well as institutions of higher learning and vocational education" that "operate[] a program or programs of scholarly research."[86] Additionally, the request "must serve a scholarly research goal of the institution."[87] Regarding the definition of a "noncommercial scientific institution," the institution must be one that is "operated solely for the purpose of conducting scientific research[,] the results of which are not intended to promote any particular product or industry."[88]

To qualify for the third category within the noncommercial classification as a representative of the news media, the person must "actively gather[] information of current interest to the public for an organization that is organized and operated to publish or broadcast news to the general public."[89] While the definition of what constitutes "news" is increasingly

---

83. See Pamela B. Rutledge, How Cambridge Analytica Mined Data for Voter Influence, Psychol. Today (Mar. 21, 2018), https://www.psychologytoday.com/us/blog/positively-media/201803/how-cambridge-analytica-mined-data-voter-influence [https://perma.cc/593N-6NMG] (tracing the use of personal data obtained online from the Obama campaign to the present).

84. See id. ("Everyone is hyper-sensitive to fake news no matter their political persuasion. Knowing that misinformation was at the root of this data gathering will make the violation seem even more egregious to many, especially given the cognitive bias that makes us attribute behaviors or intentions based on past experience.").

85. See, e.g., Isobel Asher Hamilton, Facebook Is Facing Multiple Class Action Lawsuits over the Cambridge Analytica Scandal, Bus. Insider (July 11, 2018), https://www.businessinsider.com/facebook-faces-first-class-action-cambridge-analytica-2018-7 [https://perma.cc/LCA9-CBBC]; see also Bay City News Service, Facebook Users and Shareholders File Four Lawsuits Over Data Harvesting, SFGate (Mar. 23, 2018), https://www.sfgate.com/news/bayarea/article/Facebook-Users-And-Shareholders-File-Four-12775748.php (on file with the *Columbia Law Review*).

86. See FOIA Guide, supra note 61 (quoting OMB Fee Guidelines, 52 Fed. Reg. 10,012, 10,018 (Mar. 27, 1987)) (defining and classifying different types of FOIA requests for the purpose of assessing fees).

87. Id.

88. Id.

89. Id.; see also Nat'l Sec. Archive v. U.S. Dep't of Def., 880 F.2d 1381, 1387 (D.C. Cir. 1989) ("A representative of the news media is, in essence, a person or entity that gathers

under debate, FOIA case law is instructive in teasing out an administrable definition of news media that can be employed in the scraping context.[90] For example, freelance journalists, "when they can demonstrate a solid basis for expecting the information disclosed to be published by a news organization," can be properly classified as members of the news media.[91] However, this category remains narrow by excluding "private repositories" or "middlemen" such as "information vendors" who request records for use by others.[92] This narrow definition of noncommercial actors is beneficial in the scraping context because it tightly circumscribes scraping meant to serve the public and excludes actors like Cambridge Analytica that harvest users' information under the guise of research.[93] Noncommercial scrapers comprise a far smaller share of existing scraping litigation.[94] Only three opinions of those in Sellars's overview of scraping litigation generally involve a public-interest-oriented scraper and a commercial data host.[95]

Web scraping is a popular technique in journalism and is used to serve First Amendment values of democratic self-governance, autonomy, and truth.[96] Journalists—both those with a coding specialty and without—are utilizing scraping methods and tools to uncover how collections of data

---

information of potential interest to a segment of the public, uses its editorial skills to turn the raw materials into a distinct work, and distributes that work to an audience."); Elec. Privacy Info. Ctr. v. U.S. Dep't of Def., 241 F. Supp. 2d 5, 14 (D.D.C. 2003) (explaining that the fact that an entity distributes its publication "via the Internet to subscribers' e-mail addresses does not change the [news media] analysis").

90. See FOIA Guide, supra note 61 ("Indeed, since 2000, there have been no fewer than nine district court FOIA decisions on this issue that have arisen within the D.C. Circuit, with eight involving the same plaintiff organization.").

91. Id.

92. See *Nat'l Sec. Archive*, 880 F.2d at 1387 (holding that the National Archive is not a middleman because it "uses its editorial skills to turn the raw materials into a distinct work"). "Middlemen" or "private repositories" are parties that merely make FOIA requests as an intermediary on behalf of others and do not exercise editorial discretion over what information is requested and how it is used once received. See id.

93. See supra notes 77–85 and accompanying text.

94. See Sellars, supra note 8, at 389–91 (identifying sixty-one opinions on web scraping and noting that only three involve noncommercial hosts).

95. See id. at 390–91; see also Sandvig v. Sessions, 315 F. Supp. 3d 1, 8–10 (D.D.C. 2018) (discussing scraping proposed by academics, researchers, and journalists to investigate discrimination on the internet); 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) (discussing scraping by a hacker who discovered a data vulnerability on AT&T's website and disclosed the security oversight to an online publication); VRCompliance LLC v. HomeAway, Inc., No. 1:11-cv-1088, 2011 WL 6779320, at *1 (E.D. Va. Dec. 27, 2011) (discussing scraping on behalf of resort towns to determine if online a home-rental platform evaded tax obligations).

96. See Baranetsky, supra note 12 (noting that scraping is a popular technique among journalists, including those without a coding specialty); see also Shelly Tan, Five Data Scraping Tools for Would-Be Data Journalists, Knight Lab (Mar. 20, 2014), https://knightlab.northwestern.edu/2014/03/20/five-data-scraping-tools-for-would-be-data-journalists/ [https://perma.cc/3CVR-HDGH] (sharing new scraping tools useful for journalists).

influence our everyday lives.[97] For example, ProPublica journalists conducted a computer-assisted investigation of Amazon's pricing algorithm.[98] In that report, a software program simulated a non-Prime Amazon member and found that Amazon's "algorithm pushes its own products ahead of better deals offered by others."[99] Similarly, in 2016, *Atlanta Journal-Constitution* journalists used scraping as a research tool to investigate incidents of sex abuse and misconduct among doctors across the nation.[100] In an effort to investigate doctors who were caught sexually abusing their patients yet allowed to continue practicing medicine, the reporters used scrapers to find public disciplinary orders on the websites of medical boards and regulatory agencies.[101] In contrast to the reporters' previous efforts to obtain such documents via public records requests—a tactic that yielded few results[102]—the scraping tool gathered more than 100,000 documents that were hidden in obscurity online.[103]

The work of digital journalist Julia Angwin provides another urgent example of the importance of scraping as a journalistic tool that can serve First Amendment values. Following the revelation that Russian Facebook advertisements targeted and influenced U.S. voters before the 2016 election,[104] Angwin—alongside ProPublica—built a browser plugin that "allow[ed] Facebook users to automatically send . . . the ads that are

---

97. See Baranetsky, supra note 12 (cataloguing various recent news stories made possible by scraping techniques).

98. Julia Angwin & Surya Mattu, Amazon Says It Puts Customers First. But Its Pricing Algorithm Doesn't, ProPublica (Sept. 20, 2016), https://www.propublica.org/article/amazon-says-it-puts-customers-first-but-its-pricing-algorithm-doesnt [https://perma.cc/627S-UDKF].

99. Id.; see also Julia Angwin & Surya Mattu, How We Analyzed Amazon's Shopping Algorithm, ProPublica (Sept. 20, 2016), https://www.propublica.org/article/how-we-analyzed-amazons-shopping-algorithm [https://perma.cc/M674-UBDZ] (describing the methodology for ProPublica's study).

100. See Carrie Teegardin, Danny Robbins, Jeff Ernsthausen & Ariel Hart, License to Betray, Atlanta J.-Const. (July 6, 2016), http://doctors.ajc.com/doctors_sex_abuse/?ecmp=doctorssexabuse_microsite_nav [https://perma.cc/U95Y-TYBL].

101. See Carrie Teegardin, Behind the Scenes: How the Doctors & Sex Abuse Project Came About, Atlanta J.-Const. (Dec. 17, 2016), https://www.ajc.com/news/opinion/behind-the-scenes-how-the-doctors-sex-abuse-project-came-about/UKFjNSqXoVOF9754k4wZ3M/ [https://perma.cc/V9MC-2MBP].

102. See id. ("After filing public records requests with the 64 state agencies that license or discipline doctors in every state, reporters found that those kind of databases weren't kept.").

103. See id.

104. See Issie Lapowsky, How Russian Facebook Ads Divided and Targeted US Voters Before the 2016 Election, WIRED (Apr. 16, 2018), https://www.wired.com/story/russian-facebook-ads-targeted-us-voters-before-2016-election/ [https://perma.cc/R3H9-L5ZY] (describing Russian trolls' "influence campaign" on Facebook preceding the 2016 presidential election); Nicholas Thompson & Fred Vogelstein, Inside the Two Years that Shook Facebook—And the World, WIRED (Feb. 12, 2018), https://www.wired.com/story/inside-facebook-mark-zuckerberg-2-years-of-hell/ [https://perma.cc/86D6-2RL4] (discussing the fallout at Facebook following the discovery of Russian propaganda on the platform).

displayed in their News Feeds, along with their targeting information."[105] The compilation revealed the political ads based on who was meant to see them. There is a continued appetite for tech-accountability journalism as demonstrated by Angwin's highly praised new venture *The Markup*, which aims to investigate technology's societal impacts.[106] Angwin has cited a desire for accountability as one of her main motivations behind pursuing such a project.[107] She noted that it was difficult "to know how . . . ads affected people because . . . [i]t's all ephemeral, which is a problem because if there's anything the public should be able to see and fact-check, it's political ads."[108]

Journalists are not the only ones pursuing scraping as a method of promoting accountability; academics also scrape web data for uses arguably in the public interest. For example, Harvard Business School researchers used web browser automation tools to scrape data from Airbnb host inquiry responses.[109] That study found that people with "African-American sounding names were 16 percent less likely to have their rental requests accepted compared to identical guests with distinctly white names."[110] Scraping has also been proposed as a method of information preservation. For example, web scraping can be an easy and low-cost method of preserving government documents as government publications have shifted from print to electronic formats.[111] Similarly, a group of scientists used scraping techniques to preserve information from government websites and discovered that the National Park Service had removed politically inconvenient environmental information related to efforts to reduce carbon emissions.[112]

---

105. Jeremy B. Merrill, Ally J. Levine, Ariana Tobin, Jeff Larson & Julia Angwin, Facebook Political Ad Collector, ProPublica (July 17, 2018), https://projects.propublica.org/facebook-ads/ [https://perma.cc/L824-92D4].

106. Nellie Bowles, News Site to Investigate Big Tech, Helped by Craigslist Founder, N.Y. Times (Sept. 23, 2018), https://www.nytimes.com/2018/09/23/business/media/themarkup-craig-newmark.html (on file with the *Columbia Law Review*).

107. See Jeff John Roberts, News Sites that Take on Big Tech Face Legal Peril, Fortune (Sept. 27, 2018), http://fortune.com/2018/09/27/facebook-research-censorship/ [https://perma.cc/8WHF-H6FM] (interviewing Angwin about her motivations for starting *The Markup*).

108. Id.

109. Benjamin Edelman, Michael Luca & Dan Svirsky, Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment, Am. Econ. J., Apr. 2017, at 1, 7.

110. Baranetsky, supra note 12.

111. See Rebecca Kunkel, Law Libraries and the Future of Public Access to Born-Digital Government Information, 109 Law Libr. J. 67, 76–77 (2017).

112. See About, Envtl. Data & Governance Initiative, https://envirodatagov.org/about/ [https://perma.cc/8A8J-YEVE] (last visited Sept. 14, 2019) ("EDGI formed in November 2016 out of concerns about the possibility that vital environmental data and other information from government websites might be erased or altered, as well as a general concern about the future of the Environmental Protection Agency (EPA) and other environment-

B. *Interpreting the CFAA's Access Provision*

Although the uses of web scraping are broad and varied, both prosecutors and private parties turn to the CFAA as a remedy for alleged misuse of data obtained via scraping.[113] The CFAA is an antihacking statute passed in 1986 that imposes both criminal and civil liability on a person who accesses a protected computer without proper authorization.[114] The statute lays out several categories of prohibited activity, such as accessing a computer to obtain information relating to national security,[115] financial records,[116] or government agency information;[117] obtaining information without authorization from "protected computers";[118] engaging in fraudulent activity;[119] transmitting malware with an intent to cause damage;[120] and recklessly causing damage and loss while accessing a computer without authorization.[121] The civil liability expansion is stipulated in section

---

related agencies."); see also Sarah Emerson, The National Park Service Promises to Reinstate 92 Climate Change Documents Removed from Website, Vice: Motherboard (Dec. 22, 2017), https://motherboard.vice.com/en_us/article/j5vpak/the-national-park-service-promises-toreinstate-92-climate-change-pages-removed-from-website [https://perma.cc/PVP5-E426?type=image].

113. See, e.g., United States v. Nosal, 676 F.3d 854, 858 (9th Cir. 2012) (government action); hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099, 1104 (N.D. Cal. 2017), aff'd and remanded, No. 17-16783, 2019 WL 4251889 (9th Cir. Sept. 9, 2019) (private party action); Craigslist, Inc. v. 3taps, Inc., 942 F. Supp. 2d 962, 967–68 (N.D. Cal. 2013) (same).

114. 18 U.S.C. § 1030 (2012).

115. Id. § 1030(a)(1).

116. Id. § 1030(a)(2)(A).

117. Id. § 1030(a)(2)(B).

118. Id. § 1030(a)(2)(C). The statute defines "protected computer" as either a computer used by a financial institution or the United States government, or a computer "used in or affecting interstate or foreign commerce or communication." 18 U.S.C. § 1030(e)(2)(A)–(B). Because of this sweeping definition, a protected computer is merely one that is connected to the internet. See United States v. Trotter, 478 F.3d 918, 921 (8th Cir. 2007) (holding that the Salvation Army's computers were "protected" merely because they had a connection to the internet and thus affected interstate commerce); United States v. Drew, 259 F.R.D. 449, 457–58 (C.D. Cal. 2009) (noting that the "protected computer" element is satisfied "whenever a person using a computer contacts an Internet website and reads any response from that site"); Paradigm Alliance, Inc. v. Celeritas Techs., LLC, 248 F.R.D. 598, 602 (D. Kan. 2008) ("As a practical matter, a computer providing a 'web-based' application accessible through the internet would satisfy the 'interstate communication' requirement."); see also Dep't of Justice, Prosecuting Computer Crimes 4–5 (2015), https://www.justice.gov/sites/default/files/criminal-ccips/legacy/2015/01/14/ccmanual.pdf [https://perma.cc/G3J7-6SLK] ("[I]t is enough that the computer is connected to the Internet; the statute does not require proof that the defendant also used the Internet to access the computer or used the computer to access the Internet.").

119. 18 U.S.C. § 1030(a)(4). Section 1030(a)(4) has a narrow focus on fraudulent activity and makes it a crime when a person "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." Id.

120. Id. § 1030(a)(5)(A).

121. Id. § 1030(a)(5)(B)–(C).

1030(g), which is also triggered when the offender's actions meet one of the five elements set out in sections 1030(c)(4)(A)(i)(I)–(V).[122]

The CFAA implicates all web scraping activity—regardless of its place on the scraping taxonomy—because of how courts interpret the so-called access provision inherent in each of the CFAA sections imposing liability.[123] All of the sections imposing liability—sections 1030(a)(2), 1030(a)(4), and 1030(a)(5)—include either the phrase "without authorization" or "exceeds authorized access." The CFAA defines to "exceed[] authorized access" as "to access a computer with authorization and to use such access to obtain or alter information in the computer that the accesser is not entitled so to obtain or alter."[124] However, because there is no definition of what constitutes "authorized" or "unauthorized" computer access, the term "authorization" has been defined and applied inconsistently across various district and circuit courts.[125] The CFAA "was intended to penalize hackers for breaking and entering into another person's computer."[126] Whereas an earlier version of the statute only narrowly criminalized efforts to "hack" and break into computers to obtain government information or financial information,[127] the current iteration of the statute broadly penalizes any action that can be characterized as unauthorized access.[128]

Due to the statutory ambiguity, courts are split as to how to interpret unauthorized access. Three federal courts of appeals—the First, Fifth, and Eleventh Circuits—have broadly interpreted the CFAA to include violations of a corporation's terms of use policies.[129] The sweeping definitions adopted by these circuits focus on the mindset of the computer

---

122. Id. § 1030(g).

123. See Sellars, supra note 8, at 391 ("The text of the CFAA generally does not draw distinctions based on the purpose for which one accesses a computer without authorization.").

124. 18 U.S.C. § 1030(e)(6).

125. See Myra F. Din, Note, Breaching and Entering: When Data Scraping Should Be a Federal Computer Hacking Crime, 81 Brook. L. Rev. 405, 418–26 (2015) (describing inconsistencies in courts' interpretations of the word "authorization" and potential penalties arising as a result of courts' interpretations).

126. Baranetsky, supra note 12; see also United States v. Nosal, 676 F.3d 854, 858 (9th Cir. 2012) ("Congress enacted the CFAA in 1984 primarily to address the growing problem of computer hacking, recognizing that, '[i]n intentionally trespassing into someone else's computer files, the offender obtains at the very least information as to how to break into that computer system.'" (alteration in original) (quoting S. Rep. No. 99–432, at 9 (1986) (Conf. Rep.), reprinted in 1986 U.S.C.C.A.N. 2479, 2487)).

127. See Comprehensive Crime Control Act of 1984, Pub. L. No. 98-473, 98 Stat. 2190 (codified as amended at 18 U.S.C. § 1030(a)(1)–(3)).

128. 18 U.S.C. § 1030(a)(2)(C); see also Lee Goldman, Interpreting the Computer Fraud and Abuse Act, Pitt. J. Tech. L. Pol'y, Fall 2012, at 1, 8 (describing how amendments to the CFAA have shaped the activity governed by the law).

129. See United States v. Rodriguez, 628 F.3d 1258, 1263 (11th Cir. 2010) (holding that accessing personal records for nonbusiness purposes violated the CFAA); United States v. John, 597 F.3d 263, 271 (5th Cir. 2010) (holding that exceeding the "limits placed on *the use* of information," even if access to such information is otherwise permitted, violates the

owner, looking to their intentions and relationships.[130] This means that a broad spectrum of behavior may be subject to criminal liability as long as a company lists the infraction in its terms of service. Under this interpretation, a company need only prohibit "scraping" or "data collection" in its terms of service to trigger CFAA protections.[131]

Other courts—including the Second, Fourth, and Ninth Circuits—narrowly construe the CFAA as an antihacking statute that only penalizes access if it amounts to "breaking and entering" a computer without any lawful access at all[132]—as was initially intended by Congress when the CFAA was passed.[133] Under the narrow approach, exceeding the bounds of permitted access would not be penalized under the CFAA. In 2015, the Second Circuit held in *United States v. Valle* that a narrow interpretation of the statute is "consistent with the statute's principal purpose of addressing the problem of hacking, i.e., trespass into computer systems or data."[134] Similarly, the Fourth Circuit in *WEC Carolina Energy Solutions LLC v. Miller* stated it could not "contravene Congress's intent by transforming a statute meant to target hackers into a vehicle for imputing liability to workers who access computers or information in bad faith, or who disregard a use policy."[135]

Courts are also split over the more specific issue of whether violating a website's terms of service by engaging in prohibited conduct—such as scraping—exceeds authorized access under the CFAA.[136] This debate is

---

CFAA); EF Cultural Travel BV v. Zefer Corp. 318 F.3d 58, 62 (1st Cir. 2003) ("A lack of authorization could be established by an explicit statement on the website restricting access."); see also Int'l Airport Ctrs., LLC v. Citrin, 440 F.3d 418, 420–21 (7th Cir. 2006) (holding that an employee who deleted his employer's files in violation of his employment contract had terminated the agency relationship that authorized him to access the information).

130. See Baranetsky, supra note 12 (referencing contracts, relationships, and expectations on the part of the computer owner as playing a significant role in courts' CFAA infringement analysis).

131. Id.

132. See United States v. Valle, 807 F.3d 508, 523–28 (2d Cir. 2015) (adopting a narrow interpretation because of the rule of lenity); WEC Carolina Energy Sols. LLC v. Miller, 687 F.3d 199, 206 (4th Cir. 2012) (improperly using information obtained via lawful access to company information does not violate the CFAA); United States v. Nosal, 676 F.3d 854, 863 (9th Cir. 2012) (retrieving confidential information via company user accounts and transferring it to a competitor does not violate the CFAA); see also Pulte Homes, Inc. v. Laborers' Int'l Union of N. Am., 648 F.3d 295, 304 (6th Cir. 2011) (holding that unwanted emails and phone calls "at most" exceeded authorized access, but did not constitute access without authorization).

133. See supra note 18 and accompanying text.

134. 807 F.3d at 526.

135. 687 F.3d at 207.

136. Compare Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067 (9th Cir. 2016) ("[A] violation of the terms of use of a website—without more—cannot establish liability under the CFAA."), cert. denied, 138 S. Ct. 313 (2017) (mem.), *Valle*, 807 F.3d at 528 (rejecting the government's interpretation of "exceeds authorized access" because it "makes every violation of a private computer use policy a federal crime" (internal quotation

most clearly highlighted in the Ninth Circuit case *Facebook v. Power Ventures*, for which the Supreme Court recently denied certiorari.[137] In that case, the Ninth Circuit held that a violation of a website's terms of use plus some other additional factor—which has not been specified—could ostensibly establish liability under the CFAA.[138] The "additional factor" could be as simple as refusing to comply with a cease and desist letter, which would demonstrate that the complainant had proactively revoked access and that the infringer was on notice that they exceeded the bounds of their access.[139]

The wide variety of outcomes in scraping-related litigation demonstrates that courts are uncertain of what exactly constitutes computer hacking.[140] The courts' varying interpretations of the CFAA's access provision explain the inconsistencies in outcomes and demonstrate that liability for violating a website's terms of service is an unpredictable business—particularly for researchers and journalists operating on a national scale often with limited resources. Understanding such risk of liability is key to understanding why First Amendment protection may be warranted for certain types of scraping activities, as discussed in Part II.

II. COMPETING FIRST AMENDMENT INTERESTS

As described in Part I, courts' interpretations of the CFAA access provision make scraping a legally risky activity by effectively imposing liability for breaching a website's terms of service by engaging in scraping. The language of the CFAA does not draw distinctions between the various

---

marks omitted) (quoting *Nosal*, 676 F.3d at 859)), and *WEC Carolina Energy Sols.*, 687 F.3d at 206 (noting that the unauthorized use reading "would impute liability to an employee who with commendable intentions disregards his employer's policy against downloading information to a personal computer so that he can work at home"), with EarthCam, Inc. v. OxBlue Corp., 703 F. App'x. 803, 808 & n.2 (11th Cir. 2017) (stating that "one of the lessons from [circuit precedent] may be that a person exceeds authorized access if he or she uses the access in a way that contravenes *any* policy or term of use governing the computer in question," and noting the dissenting views of other circuits), CollegeSource, Inc. v. AcademyOne, Inc., 597 F. App'x. 116, 130 (3d Cir. 2015) (suggesting that defendants can be prosecuted under the CFAA if they "breach[ed] any technological barrier or contractual term of use"), and EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58, 62 (1st Cir. 2003) ("A lack of authorization could be established by an explicit statement on the website restricting access. . . . Many webpages contain lengthy limiting conditions, including limitations on the use of scrapers."). For a summary of the existing circuit splits surrounding the CFAA, see Sandvig v. Sessions, 315 F. Supp. 3d 1, 22–23 (D.D.C. 2018) (noting that the D.C. Circuit has never taken a side of the circuit split over "whether violating a website's ToS exceeds authorized access for the purposes of the CFAA").

137. 844 F.3d at 1058, cert. denied, 138 S. Ct. 313 (2017) (mem.).

138. See id. at 1067–69.

139. See id.

140. See Din, supra note 125, at 407 (suggesting that the definition of what constitutes computer hacking is "unsettled" because of ever-evolving methods of internet communication and computer hacking since the CFAA was first passed).

types of scraping that occur in society.[141] While some forms of scraping may be more innocuous than others, the CFAA crudely lumps together very different forms of scraping with very different motivations and implications for social values.

This Part argues that there are competing First Amendment interests at stake in data-scraping practices as implicated by the CFAA access provision. On one hand, right-to-record jurisprudence[142] suggests that researchers should be able to use data-scraping techniques to gather information. On the other hand, there is arguably a First Amendment interest in privacy. A right to scraping could present scary implications for consumers—and potentially have a chilling effect on their speech—if data they entrust to social media companies are widely accessible to third parties. This Part assesses the First Amendment interests at stake and paves the way for a discussion in Part III of how the CFAA can be reformed to impose liability only after balancing these competing First Amendment interests.

A. First Amendment Interest in the Right to Record

This section presents a claim that online scraping designed to serve the public interest—that is, scraping that falls within the second classification of this Note's taxonomy[143]—is affected with a First Amendment interest and merits protection as a kind of newsgathering or information-gathering conduct.

1. *Right-to-Record Jurisprudence.* — The First Amendment protects "freedom of speech, or of the press."[144] Many activities protected by this clause of the First Amendment are not explicitly enumerated.[145] This ambiguity regarding what activity is potentially entitled to First Amendment protection has been termed the "coverage problem" by First Amendment scholars.[146] Because web scraping is a form of conduct instead of a purely verbal act, this section will examine First Amendment protections for "conduct incidental to speech."[147]

---

141. See supra note 123 and accompanying text.

142. Right-to-record jurisprudence upholds the right to record activity, traditionally via video capture, in public spaces. This theory first emerged in the form of the right to record police activity. See Patel, supra note 46, at 1501 & n.189 (noting that circuits that have considered the issue have unanimously held that citizens have a right to record police activity). This theory has been expanded to include other forms of activity in both public and private spaces. See infra section II.A.1.

143. See supra section I.A.2.b.

144. U.S. Const. amend. I.

145. See Patel, supra note 46, at 1484 & n.72 (describing how freedom of expression and campaign finance protections have developed as rights implied by the First Amendment).

146. See Justin Marceau & Alan K. Chen, Free Speech and Democracy in the Video Age, 116 Colum. L. Rev. 991, 998 & n.28 (2016) (collecting scholarship exploring whether various types of activities qualify for First Amendment protection).

147. Patel, supra note 46, at 1485–88 (describing First Amendment protections for "conduct incidental to speech").

The doctrinal line of protection most analogous to scraping is conduct "incidental to, or in preparation for, speech."[148] In some instances, conduct in preparation for speech has also been described as a form of newsgathering or information gathering because journalists and activists are "seek[ing] access to information for the purpose of subsequently engaging in speech."[149] In cases falling under the First Amendment incidental conduct doctrine, "the Court has protected the *means* of various kinds of speech."[150]

Professors Justin Marceau and Alan Chen have examined where the line for protection of conduct incidental to speech should be drawn; they posit a spectrum theory that suggests that expressive activity occurs along a continuum of action that is a precursor to speech.[151] One end of the spectrum represents activity necessary to engage in communication and the other end of the spectrum represents "directly communicative element[s] of the expressive process—shouting through a megaphone, exhibiting a painting, displaying a video."[152] Marceau and Chen's spectrum theory has been relied upon in "right-to-record" jurisprudence.[153] They argue that video recording is a form of expression covered by the First Amendment because it falls on the spectrum of conduct essential to speech.[154] Because video recording is intertwined with expression, "its regulation must comply with constitutional safeguards for speech."[155]

The right to record police activity is one area that highlights First Amendment protections afforded to recording activity. Circuits that have considered the issue have all held that the First Amendment protects the right of

---

148. Id. at 1485.

149. Id.

150. Id. It is worth noting that the Supreme Court has not yet established a framework for when newsgathering actions merit protection. See id. at 1500 & n.185 (noting that the Supreme Court has not established a definitive framework for right-of-access-to-information cases (citing S.H.A.R.K. v. Metro Parks Serving Summit Cty., 499 F.3d 553, 560 (6th Cir. 2007))).

151. Marceau & Chen, supra note 146, at 1019 ("[E]xpressive activity typically takes place along a continuum of actions that include not only direct expression but also much of the conduct that is a necessary precursor to speech.").

152. Id.

153. See Patel, supra note 46, at 1500–01.

154. Professor Jane Bambauer has made a similar argument that the First Amendment should protect the creation of knowledge in the data context. See Jane Bambauer, Is Data Speech?, 66 Stan. L. Rev. 57, 63–64, 86–105 (2014) (arguing that data qualifies as protected speech because it serves the purpose of knowledge creation). Professor Bambauer's argument that it is possible—and necessary—to break down the distinction between conveying and collecting information is directly relevant in the recording context. For a comparison between Bambauer's argument and the right-to-record analogy, see Marceau & Chen, supra note 146, at 1019–20.

155. Marceau & Chen, supra note 146, at 997.

individuals to record official police activity.[156] These courts have reasoned that the First Amendment must protect both the initial act of creating photos, videos, and recordings and the final speech act.[157] In so holding, courts have recognized that recording is a necessary step toward expressive activity that cannot be disentangled from the final product itself.[158] This inextricable link between conduct preparing speech and speech itself has also been demonstrated in the context of so-called ag-gag laws—laws that penalize whistleblowers who investigate "the day-to-day activities of industrial farms, including the recording, possession or distribution of photos, video and/or audio at a farm."[159] In contemplating the constitutionality of an ag-gag law, one court noted that authorities would not be able to punish surreptitious recording by activists unless such recording is published because authorities would be otherwise unaware that such recording occurred.[160] Because enforcement for recording and enforcement for publishing that recording may be coextensive, the argument can be made that the act of publishing is essentially the act of recording. Additionally, the accountability function served by recordings has been a consideration at the forefront of courts' First Amendment calculus.[161] One court

156. Patel, supra note 46, at 1501 ("[T]he First, Third, Fifth, Ninth, and Eleventh Circuits have held that citizens have the right to record police activity, while the Seventh Circuit has held that there is a right to intercept communications of public officials engaged in their public duties."). The cases relied upon by Patel in her discussion of existing right-to-record jurisprudence provide further elaboration on the right to record. See Fields v. City of Philadelphia, 862 F.3d 353, 356 (3d Cir. 2017); Turner v. Lieutenant Driver, 848 F.3d 678, 690 (5th Cir. 2017); Gericke v. Begin, 753 F.3d 1, 7 (1st Cir. 2014); ACLU of Ill. v. Alvarez, 679 F.3d 583, 597–600 (7th Cir. 2012); Smith v. City of Cumming, 212 F.3d 1332, 1333 (11th Cir. 2000); Fordyce v. City of Seattle, 55 F.3d 436, 439 (9th Cir. 1995).

157. E.g., *Fields*, 862 F.3d at 358 ("There is no practical difference between allowing police to prevent people from taking recordings and actually banning the possession or distribution of them."); see also *Alvarez*, 679 F.3d at 597–600 (viewing prohibitions on recording as "necessarily limit[ing] the information that might later be published or broadcast—whether to the general public or to a single family member or friend—and thus burden[ing] First Amendment rights").

158. See, e.g., Animal Legal Def. Fund v. Wasden, 878 F.3d 1184, 1203 (9th Cir. 2018) (suggesting that disaggregating the process of video creation from the video or audio recording itself "defies common sense"); *Turner*, 848 F.3d at 688–89 (suggesting that protecting film itself also protects the process of creating the film); Glik v. Cunniffe, 655 F.3d 78, 82 (1st Cir. 2011) ("[T]he First Amendment's aegis . . . encompasses a range of conduct related to the gathering and dissemination of information.").

159. What Is Ag-Gag Legislation?, ASPCA, https://www.aspca.org/animal-protection/public-policy/what-ag-gag-legislation [https://perma.cc/KC79-Z7G2] (last visited Sept. 15, 2019).

160. See Animal Legal Def. Fund v. Otter, 44 F. Supp. 3d 1009, 1023 (D. Idaho 2014) ("[A]n undercover investigator who never publishes a video after surreptitiously filming a facility's operations will likely never be punished for the filming because, in most cases, authorities will not become aware of a violation of the statute until a video is published.").

161. See Patel, supra note 46, at 1502 ("Another weighty consideration pushing . . . courts to recognize a 'right to record' was the notion that such recordings may assist in unveiling and correcting police misconduct.").

noted that videos recorded by civilians have helped jumpstart civil rights investigations and trigger policing reform.[162]

The right to record is not limited to recording that occurs in the public sphere.[163] Courts have upheld a right to record on private property in several different contexts, including the recording of abuse at agricultural facilities[164] and an ABC News investigation of unsanitary food practices at Food Lion stores.[165] In fact, the private nature of these facilities may strengthen the argument for allowing recording to take place: Without video, "it would be impossible to investigate occurrences hidden behind an 'enforced wall of secrecy.'"[166] Although collapsing the public–private distinction has not been widely adopted by courts in this context, there is a compelling argument that recording—regardless of location—serves fundamental First Amendment values of revealing the truth.[167]

2. *Right to Record and Scraping*. — The right to record is directly analogous and relevant to the acts of scraping that fall within the classification outlined in section I.A.2.b—noncommercial web scraping for research and journalistic purposes.[168] Under Marceau and Chen's spectrum

---

162. See *Fields*, 862 F.3d at 359–60 (noting that bystander videos have helped with civil rights investigations and helped address police misconduct); see also Gentile v. State Bar of Nev., 501 U.S. 1030, 1034–35 (1991) (recognizing a core First Amendment interest in "the dissemination of information relating to alleged governmental misconduct"); Press-Enter. Co. v. Superior Court, 478 U.S. 1, 8 (1986) (noting that in many situations, government "operate[s] best under public scrutiny"); *Turner*, 848 F.3d at 689 ("Filming the police contributes to the public's ability to hold the police accountable, ensure that police officers are not abusing their power, and make informed decisions about police policy."). For further discussion of the potential benefits to democracy of recording policy activity, see Patel, supra note 46, at 1502 & n.195 (discussing how civilian videos can "fill[] the gaps when recordings of police conduct are otherwise unavailable or withheld").

163. See Marceau & Chen, supra note 146, at 1023 ("[R]ecording[s] . . . [made] on private property, just as recordings made in public, advance the fundamental free speech values of promoting democracy and facilitating the search for truth.").

164. See supra notes 159–160 and accompanying text; see also Animal Legal Def. Fund v. Wasden, 878 F.3d 1184, 1203 (9th Cir. 2018) (striking down an Idaho law that criminalized recording agricultural production facilities without the consent of the owner because it violated the First Amendment); Animal Legal Def. Fund v. Herbert, 263 F. Supp. 3d 1193, 1208 (D. Utah 2017) (finding "no support in the case law" that private agricultural facilities do not enjoy First Amendment protection).

165. See Food Lion, Inc. v. Capital Cities/ABC, Inc., 194 F.3d 505, 510–11 (4th Cir. 1999).

166. Marceau & Chen, supra note 146, at 1024 (quoting United States v. Biasucci, 786 F.2d 504, 511 (2d Cir. 1986)).

167. See id. ("[S]hutting down [video recording's] production interferes with expression and also impedes the creation of knowledge and information. It simply cannot be the rule that the state may ban nondisruptive recording of nonintimate matters just because they occur on private property." (footnote omitted)); see also Patel, supra note 46, at 1505 ("[T]he idea that speech acts on private facilities enjoy no First Amendment protection finds 'no support in the case law.'" (quoting *Herbert*, 263 F. Supp. 3d at 1208)).

168. See Patel, supra note 46, at 1500 (applying Marceau and Chen's spectrum theory to argue that civil rights researchers' testing methods—which include scraping—merit First

theory, there is little to no attenuation between the speech activity (the publication of scraped data results) and the conduct meriting protection (the scraping itself).[169] Other legal scholars have made similar comparisons between speech and conduct undertaken in preparation for speech, asking, "In the context of an exposé, what is more directly antecedent to its writing and subsequent publication than the investigation that yielded its content?"[170]

Although litigation involving scraping done in the public interest is limited,[171] courts are beginning to employ right-to-record analogies in scraping litigation. The District Court for the District of Columbia was likely the first court to recognize the right-to-record theory within the scraping context in *Sandvig v. Sessions*.[172] In that case, academics, researchers, and journalists brought a constitutional challenge to the CFAA, arguing that the CFAA violated the First Amendment because it chilled their constitutionally protected access to research and journalism by imposing liability for using digital research tools that violate a website's terms of service.[173] The plaintiffs sought to use automated web browsing tools to investigate online discrimination.[174] Such tools included the creation of bots, "automated agents that will each browse the Internet and interact with websites as a human user might,"[175] and the use of "scraping to record the properties that each bot sees on the real estate sites."[176] Similarly, other researchers intended to use similar tactics in the employment context.[177] Finally, the journalist and researcher plaintiffs sought to examine the discriminatory effects of algorithms.[178] However, they faced the risk of liability under the CFAA since "to knowingly violate some of [a website's] terms . . . could get one thrown in jail."[179]

While the court did not strike down the CFAA as unconstitutional, Judge John D. Bates did recognize a First Amendment interest in preventing CFAA liability from applying to journalists and researchers undertaking research projects that serve the public interest.[180] The court held that

---

Amendment protection as information-gathering conduct that falls close enough to speech).

169. See Marceau & Chen, supra note 146, at 1011–23 (arguing that there is such a close link between video recording and speech that recording can be considered either a form of speech or conduct preparatory to speech).

170. Patel, supra note 46, at 1500.

171. See supra note 95 and accompanying text.

172. 315 F. Supp. 3d 1, 15–16 (D.D.C. 2018).

173. See id. at 8–10.

174. See id.

175. Id. at 9.

176. Id.

177. See id. at 9–10.

178. Id.

179. Id. at 8.

180. See id. at 13–17.

the researchers' claim that their planned acts of scraping to record information from private websites for research purposes were arguably affected with a First Amendment interest was plausible.[181] Judge Bates noted that the internet is fertile ground for First Amendment activity and thus the internet merits "special First Amendment protection."[182]

A crucial part of the *Sandvig* decision, however, was Judge Bates's holding that whether the websites scraped are privately run does "not change the calculus"[183] because the public–private distinction that is ordinarily applied in the real world does not easily map on to regulation of the internet.[184] Judge Bates noted that the "public" internet is "too heavily suffused with First Amendment activity" and the online equivalent of "private spaces are too blurred with expressive spaces," which makes drawing a line based on the public–private distinction used in physical space problematic.[185] Take, for example, the food truck hypothetical posed by the *Sandvig* court.[186] A traditional privately owned restaurant can remove a disruptive customer from its private property without encountering any First Amendment barriers.[187] However, a food truck parked in a public park might confront a First Amendment barrier if seeking to remove the same customer standing in the vicinity of their truck because the sidewalk is a public space.[188] Although the truck is privately owned and customers interact with the truck for the "private purpose of buying a meal," the disruptive customer is still in a public forum—the sidewalk—and "her speech remains protected even when she interacts with a private business located within that forum."[189] The internet is analogous to the food truck: Some websites simultaneously exist as private, commercial spaces while also facilitating public interaction and activity.[190]

---

181. See id.

182. See id. at 11–12 ("Only last Term, the Supreme Court emphatically declared the Internet a primary location for First Amendment activity: 'While in the past there may have been difficulty in identifying the most important places (in a spatial sense) for the exchange of views, today the answer is clear. It is cyberspace . . . .'" (quoting Packingham v. North Carolina, 137 S. Ct. 1730, 1735 (2017))).

183. Id. at 12.

184. See id. at 12–13 ("Regulation of the Internet presents serious line-drawing problems that the public/private distinction in physical space does not.").

185. Id. at 13.

186. Id. at 12.

187. Id.

188. Id. ("Yet if a customer standing on a public sidewalk tastes her food and then yells at those in line behind her that they should avail themselves of the myriad other culinary options nearby, the truck could not call the police to arrest her for her comments.").

189. Id.

190. Id. at 13 (describing how Amazon engages in "private, commercial activity" while also encouraging protected First Amendment activity).

Still, Judge Bates did not entirely collapse the public–private distinction. Instead, he noted that the distinguishing factor is whether "the owners of the information at issue have taken real steps to limit who can access" the information.[191] For example, in the food truck hypothetical, a customer who breaks into the locked food truck parked on public property to steal confidential culinary information would not be protected by the First Amendment because they "circumvent[ed] barriers that demarcate private areas."[192] However, owners cannot simply place "contractual conditions on accounts that anyone can create . . . [to] remove a website from the First Amendment protections of the public Internet."[193] Rather, code-based barriers that limit access to information may "remove those protected portions of a site from the public forum."[194] Finally, and most significantly, Judge Bates opined that scraping "plausibly falls within the ambit of the First Amendment" based on a right-to-record analogy.[195] Defining scraping as merely a technological advancement that facilitates the recording of information, Judge Bates found that 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."[196]

The D.C. District Court's opinion in *Sandvig* demonstrates that web scraping that serves a higher public interest should be considered constitutionally protected activity under the First Amendment.[197] Scraping on platforms that play an important role in today's speech infrastructure and

---

191. Id.

192. Id. at 14 ("This is true even if the customer claimed she was doing so in order to broadcast to the world the truck's substandard ingredients and ill-conceived recipes.").

193. Id. at 13.

194. Id. For further discussion of the implications of this language on potential reform efforts, see infra notes 277–284 and accompanying text.

195. *Sandvig*, 315 F. Supp. 3d at 15 ("The First Amendment goes beyond protection of the press and the self-expression of individuals to prohibit government from limiting the stock of information from which members of the public may draw." (internal quotation marks omitted) (quoting First Nat'l Bank of Boston v. Bellotti, 435 U.S. 765, 783 (1978))).

196. Id. at 16 ("That plaintiffs wish to scrape data from websites rather than manually record information does not change the analysis."). The court also emphasized that the information's location in a public forum only enhances the plaintiffs' ability to scrape. See id. ("[T]he 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."). However, that the information was in a public forum was not the dispositive factor in Judge Bates's analysis. See supra notes 183–184 and accompanying text.

197. See, e.g., Patel, supra note 46, at 1504 (arguing for First Amendment protections of civil rights "sock puppet" and "scraping" audits because not allowing such critical research methods "undermines the ability to monitor companies transacting online").

shape public discourse—such as Facebook—especially merit First Amendment protection.[198] As the *Sandvig* court cautioned, First Amendment protection should not be afforded to the entirety of the internet; this approach would "gloss[] over the dual public and private nature of digital arenas."[199] However, prosecuting individuals for scraping information that is available to the general public may infringe on the scrapers' First Amendment rights—regardless of whether the website is publicly or privately owned.[200]

B. First Amendment Interests that Run Counter to a Right to Scrape

There is a robust right to scrape on quasi-public forums that have a structural role of supporting discourse in line with First Amendment values of self-governance, individual autonomy, and the pursuit of truth.[201] But not all forms of data scraping serve First Amendment values of accountability and self-governance.[202] In fact, there are privacy interests—arguably First Amendment interests—that run counter to permitting scraping of all forms. For example, as a result of the Cambridge Analytica scandal, it is possible that data scraping can be viewed as such an egregious invasion of privacy that users' First Amendment activity on online platforms would be chilled.[203] This section highlights First Amendment interests at the other end of the scraping spectrum by considering privacy interests that suggest that a right to scrape should be confined within the boundaries of a narrow exception.

First Amendment theory has historically been concerned with protecting the act of speaking from interference and censorship;[204] privacy of information and speech is often viewed as a hostile value and as something to be balanced with free speech as a countervailing force.[205] However, pri-

198. See *Sandvig*, 315 F. Supp. 3d at 13 (describing Facebook as "a quintessential site for protected First Amendment activity").

199. Id. (internal quotation marks omitted) (quoting The Supreme Court, 2016 Term—Leading Cases, 131 Harv. L. Rev. 233, 238 (2017)).

200. See id. at 17–18.

201. See supra notes 10–14 and accompanying text.

202. See supra section I.A.2.a.

203. See Rutledge, supra note 83 (noting the public backlash in the wake of Cambridge Analytica).

204. See Jack M. Balkin, Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society, 79 N.Y.U. L. Rev. 1, 50 (2004) (noting that free speech rights largely came from the judiciary, not public policy, and were recognized to prevent censorship or other types of government regulation).

205. See Neil M. Richards, Reconciling Data Privacy and the First Amendment, 52 UCLA L. Rev. 1149, 1161–63 (2005) (collecting examples of First Amendment scholarly critiques of data privacy).

vacy can also advance First Amendment free speech values of self-government, autonomy, and cultural freedom by creating spaces free from government surveillance and intervention.[206]

Housing a privacy interest in one specific place in the Constitution has been a persistent challenge for judges and legal scholars alike.[207] This effort has been complicated by the fact that privacy is a continuously evolving concept.[208] In the seminal reproductive rights case *Griswold v. Connecticut*, the Supreme Court suggested that the First Amendment can be used to protect privacy.[209] In that case, Justice Douglas noted that a right of privacy protecting the intimate relations of married couples is implied by the sum of the First, Third, Fourth, Fifth, and Ninth Amendments and that protected activities are "penumbras" that are not specifically enumerated in the Constitution.[210] This holding has been construed as narrowly applying to only bodily conduct,[211] likely because the *Griswold* analysis "offers no generalizable definition of the right it is used to protect."[212] However, because *Griswold* provides no direction, limiting privacy to reproductive rights and intimate relations is a matter of custom, not law.[213]

Though *Griswold* focused on reproductive rights, classic First Amendment theory similarly suggests that privacy is key to First Amendment values of autonomy, thought formation, and self-governance. Shiffrin, for example, posited that the First Amendment protects individual autonomy; it is not solely one's interactions that are integral to free speech but rather one's autonomy and ability to reflect and think freely.[214] Similarly, Meiklejohn

---

206. See, e.g., Alex Abdo, Why Rely on the Fourth Amendment to Do the Work of the First?, 127 Yale L.J. Forum 444, 444, 446–47 (2017), https://www.yalelawjournal.org/pdf/Abdo_5czbvbj9.pdf [https://perma.cc/F2SH-TQU8] [hereinafter Abdo, Why Rely] (arguing that government surveillance poses a threat to First Amendment freedoms because dissent "requires privacy and often confidential association to flourish").

207. See Note, Privacy in the First Amendment, 82 Yale L.J. 1462, 1475 (1973) ("The Supreme Court has declared that the Constitution protects a right to privacy, but the supporting analysis offers no hint as to how that protected privacy might be defined. . . . [T]he rights so defined are far too narrow and qualified to serve as a satisfactory 'right to privacy.'").

208. Jon L. Mills, Privacy: The Lost Right 21 (2008) ("As society changes, so too does its reasonable expectation of privacy . . . .").

209. 381 U.S. 479, 483 (1965) ("In other words, the First Amendment has a penumbra where privacy is protected from governmental intrusion.").

210. See id. at 482–86.

211. See, e.g., Roe v. Wade, 410 U.S. 113, 152–53 (1973) (confining the right to privacy discussion articulated in *Griswold* to pregnancy terminations).

212. Note, supra note 207, at 1476.

213. See id. ("Confining the right to privacy inside that area is sanctioned by custom, but not by anything in the reasoning of the decisions.").

214. See Shiffrin, supra note 13, at 303–04 (drawing from *Griswold* to argue that the First Amendment provides a justification for intimate association founded in the ability of individuals to forge meaningful connections through free expression including intimate associations).

has argued that the First Amendment is designed to protect an integrated system of freedom.[215] Under Meiklejohn's theory, free speech is protected so that individuals can make informed decisions about matters of government.[216] Somewhat in contrast with Meiklejohn, but still in support of the idea that privacy can support the First Amendment, is the cultural democracy theory suggested by legal scholar Jack Balkin.[217] Under this theory, freedom of expression is crucial to the development of cultural power; "people should have the right to participate in the forms of meaning-making that shape who they are."[218] Thus, internet activity that harms the free "global exchange of information, art, opinion, and ideas" can conflict with the First Amendment.[219]

A two-step speech process emerges from these various First Amendment theories: First, there is the transmission of information to the listener, and second, the listener must be able to freely process that information without interference.[220] Actions that forestall the second step—potentially by breaching the individual's private zone of self-governance and risking exposure of their personal information—could be violative of the First Amendment. For example, in *Lamont v. Postmaster General*, the Supreme Court struck down a statute providing that the Post Office deliver mail deemed "communist political propaganda" only upon written request in advance from the recipient.[221] The law was deemed unconstitutional because it required an official act—returning the reply card affirming desire to receive mail—that abridged the recipients' First Amendment rights.[222] This is relevant in the scraping context because scraping, particularly on social media platforms, can result in the revelation of individuals' beliefs.[223]

Similarly, courts have recognized that the right to speak anonymously and pseudonymously is part of the First Amendment right to free speech. In *McIntyre v. Ohio Elections Commission*, the Supreme Court held that "an

---

215. See Meiklejohn, First Amendment Absolute, supra note 11, at 255 ("The First Amendment does not protect a 'freedom to speak.' It protects the freedom of those activities of thought and communication by which we 'govern.' It is concerned, not with a private right, but with a public power, a governmental responsibility.").

216. Note, supra note 207, at 1464 (arguing that a system of free expression requires "two separate stages," including "[t]ransmission of information from speaker to listener" and "the application of that information . . . to the individual decisions of self-governance," both of which are necessary "to achieve . . . free individual choice by each citizen").

217. Jack M. Balkin, Cultural Democracy and the First Amendment, 110 Nw. L. Rev. 1053 (2016).

218. Id. at 1061.

219. See id. at 1093–95.

220. See supra note 216.

221. 381 U.S. 301, 305 (1965).

222. See id.

223. See Note, supra note 207, at 1466 ("The second stage of the system, on the other hand, takes place entirely within the mind of the individual . . . . The insulation of the individual against the chilling of his self-governing decisions is thus his ability to prevent public knowledge about himself.").

author's decision to remain anonymous, like other decisions concerning omissions or additions to the content of a publication, is an aspect of the freedom of speech protected by the First Amendment."[224] Because "identification of the author against her will" can "reveal[] unmistakably the content of her thoughts on a controversial issue," such forced identification can be "particularly intrusive."[225] Additionally, in *NAACP v. Alabama*, the Supreme Court held that state action which may have the effect of curtailing the freedom to associate is subject to the strictest scrutiny, regardless of "whether the beliefs sought to be advanced by association pertain to political, economic, religious or cultural matters."[226] It further held that "compelled disclosure of affiliation with groups engaged in advocacy may constitute as effective a restraint on freedom of association as [other] forms of governmental action" and therefore fall under the First Amendment's prohibitions.[227] Taken together, these two decisions suggest that a person who feels that they aren't guaranteed anonymous speech or freedom to associate in private may self-govern differently than they would in a private sphere. Therefore, the First Amendment must protect privacy in order to maintain a democratic system of free expression.[228]

The concept of privacy as a First Amendment interest is further evinced by Professor Neil M. Richards's notion of "intellectual privacy" which recognizes that the protection of one's intellectual activities—the development of ideas and beliefs in solitary contemplation or collaboration—is crucial to free expression.[229] Intellectual privacy allows individuals to "develop ideas and beliefs away from the unwanted gaze or interference of others."[230] Exposure of one's ideas and thoughts, via surveillance or other methods, can interfere with and skew the way people think and, therefore, speak.[231] Increasingly, First Amendment advocates have called on courts to evaluate infringements on privacy through the lens of the First Amendment, instead of solely the Fourth Amendment, since thought formulation

---

224. 514 U.S. 334, 342 (1995).

225. Id. at 355.

226. 357 U.S. 449, 460–61 (1958).

227. Id. at 462; see also Gibson v. Fla. Legislative Investigation Comm., 372 U.S. 539, 569 (1963) (Douglas, J., concurring) ("The right of association has become a part of the bundle of rights protected by the First Amendment . . . and the need for a pervasive right of privacy against government intrusion has been recognized, though not always given the recognition it deserves.").

228. See Note, supra note 207, at 1467–68.

229. See Neil M. Richards, Intellectual Privacy, 87 Tex. L. Rev. 387, 403–07 (2008) [hereinafter Richards, Intellectual Privacy] ("Intellectual privacy is the ability, whether protected by law or social circumstances, to develop ideas and beliefs away from the unwanted gaze or interference of others.").

230. Id. at 389.

231. See id. ("The ability to freely make up our minds and to develop new ideas thus depends upon a substantial measure of intellectual privacy. In this way, intellectual privacy is a cornerstone of meaningful First Amendment liberties.").

and expression of dissent are threatened by intrusions on privacy.[232] This autonomy-based argument for privacy has emerged in the cybersecurity space, with some viewing digital data security as a prerequisite to personal autonomy, trust, and privacy online.[233]

In sum, legal scholars have described multiple rationales and understandings of privacy. Professor Daniel Solove, for example, has discerned at least six different understandings of privacy, including the right to be let alone, control over personal information, and intimacy.[234] Professor Kendall Thomas has argued that the freedoms to associate and make autonomous decisions fall within a "zonal, relational[,] and decisional" conception of privacy.[235] In fact, this autonomy rationale for privacy has already emerged in the cybersecurity space: Professor David Pozen has argued that privacy theory has taken a "pluralistic turn" to form a "web of overlapping conceptions, dimensions, and values," imbuing privacy with a variety of related meanings.[236]

Scraping can pose a threat to this multifaced concept of privacy—particularly intellectual privacy—because it allows anyone to get access to personal data, such as that made available on social media websites. Even if this information is publicly available, aggregation of such material can pose a unique threat to one's autonomy because it "involves the combination of data in new, potentially unanticipated ways to reveal facts about a person that are not readily known."[237] Professor Solove has posited that such aggregation of publicly available information constitutes "decisional interference," that is, the "government's incursion into the data subject's

---

232. See Abdo, Why Rely, supra note 206, at 448 ("Dissent requires breathing space: to formulate dissenting ideas, to test and debate those ideas with close associates, to expand the association into a movement, and finally to air grievances publicly, to convince fellow citizens, and to effect political change.").

233. See David E. Pozen, Privacy-Privacy Tradeoffs, 83 U. Chi. L. Rev. 221, 235 (2016) [hereinafter Pozen, Privacy-Privacy Tradeoffs] (describing rationales for government surveillance to safeguard Americans' privacy against external threats such as foreign governments, hackers, and criminals).

234. Daniel J. Solove, Understanding Privacy 12–13 (2008).

235. Kendall Thomas, Beyond the Privacy Principle, 92 Colum. L. Rev. 1431, 1443–48 (1992).

236. Pozen, Privacy-Privacy Tradeoffs, supra note 233, at 225–28 ("In recent years, many privacy theorists have made what we might call a pluralistic turn: rejecting approaches to privacy that strive to identify its essence or its core characteristics and settling, instead, 'on an understanding of privacy as an umbrella term that encompasses a variety of related meanings.'" (quoting Neil Richards, Intellectual Privacy: Rethinking Civil Liberties in the Digital Age 9 (2015))).

237. Daniel J. Solove, A Taxonomy of Privacy, 154 U. Pa. L. Rev. 477, 506–11 (2006) ("People give out bits of information in different settings, only revealing a small part of themselves in each context. . . . When these pieces are consolidated together, however, the aggregator acquires much greater knowledge about the person's life.").

decisions regarding her private affairs."[238] The special dangers of aggregated data have been articulated in Fourth Amendment jurisprudence. In *Jones v. Carpenter*, the Supreme Court held that attaching a GPS tracking device to a vehicle and subsequent use of that device to monitor the vehicle's movements on public streets constituted a search under the Fourth Amendment because it violated a person's reasonable expectation of privacy.[239] However, Justice Sotomayor's concurrence emphasized that although a GPS device tracks publicly available information—a person's location—the compilation of this information can reveal a "comprehensive record of a person's public movements that reflects a wealth of detail about her familial, political, professional, religious, and sexual associations."[240] Collection of publicly available information online is similar to collection of data on a person's location; individually, discrete pieces of public information are harmless, but taken together, a comprehensive catalogue of public information reveals private aspects of identity that may threaten the constitutionally protected right to privacy.[241]

The threat to autonomy is enhanced because scraping increases aggregation possibilities with technological efficiency. The public's concern for privacy amid technological advances has started to be raised in the social media space. For example, the Ninth Circuit recently held that users can sue Facebook for collecting and using their facial images because "advances in technology can increase the potential for unreasonable intrusions into personal privacy."[242] In the scraping context, the public concern over such a risk is illustrated in the plethora of data privacy lawsuits that have been filed following the Cambridge Analytica scandal in which plaintiffs have expressed their privacy and autonomy concerns.[243]

---

238. Id. at 491.

239. 565 U.S. 400, 404 (2012).

240. Id. at 415 (Sotomayor, J., concurring).

241. Id. at 415–18 ("[I]t may be necessary to reconsider the premise that an individual has no reasonable expectation of privacy in information voluntarily disclosed to third parties. This approach is ill suited to the digital age, in which people reveal a great deal of information about themselves . . . [while] carrying out mundane tasks."). The Supreme Court reached a similar conclusion highlighting the dangers of aggregated data of publicly available information in *Carpenter v. United States*, in which the Court held that a person maintains a legitimate expectation of privacy in the record of their movements captured through cell phone site location data. 138 S. Ct. 2206, 2217 (2018) ("Mapping a cell phone's location over the course of 127 days provides an all-encompassing record of the holder's whereabouts. As with GPS information, the time-stamped data provides an intimate window into a person's life . . . .").

242. Patel v. Facebook, Inc., 932 F.3d 1264, 1272 (9th Cir. 2019).

243. See, e.g., Complaint at 30, Malskoff v. Facebook Inc., No. 3:18-cv-03393 (N.D. Cal. Mar. 27, 2018); Class Action Complaint at 3, O'Kelly v. Facebook Inc., No. 3:18-cv-01915 (N.D. Cal. Mar. 28, 2018); Class Action Complaint at 14, Gennock v. Facebook Inc., No. 3:18-cv-01891 (N.D. Cal. Mar. 27, 2018); Class Action Complaint at 17, Rubin v. Facebook Inc., No. 3:18-cv-01852 (N.D. Cal. Mar. 26, 2018).

Overall, there is a First Amendment harm implicated if scraping results in a chilling effect on social media speech because users are concerned for their privacy. Thus, there are competing First Amendment interests on either side of the scraping debate that must be considered in reforming the CFAA.

III. COURSE CORRECTION: CONTEMPLATING CFAA REFORM

Recent studies suggest that the sheer amount of available data in the digital universe has exploded over the past few years with 2.5 quintillion bytes of data created "each day at our current pace."[244] This shift in the amount of available data has led to massive changes in the news industry[245] with data journalism playing a driving force in newsrooms around the country.[246] As the amount of data increases and tech companies play a bigger role in everyday life, data present the opportunity for journalists to go beyond simply being the first to report the news to being the ones to tell the public what a certain development might mean.[247] It is therefore increasingly important that journalists and researchers can access information crucial to maintaining an informed public. However, corporations and government entities—which control much of this vital information—are exercising stricter controls over data.[248] Imposing liability on journalists and researchers engaging in data research via scraping techniques is one such method of control.[249] Although to date no journalists have been sued or prosecuted under the CFAA,[250] the looming threat of civil and

244. See Marr, supra note 23.

245. See Baranetsky, supra note 12 ("Journalists are quickly learning how to obtain troves of data through electronic leaks, drones, and cutting-edge computer programs that sometimes require little more than the click of a button to access information.").

246. According to a 2017 Google News Lab study, forty-two percent of reporters said they use data more than twice per week in their reporting and just over fifty percent of news organizations have a staff data journalist. See Simon Rogers, Jonathan Schwabish & Danielle Bowers, Google News Lab, Data Journalism in 2017: The Current State and Challenges Facing the Field Today 10–12 (2017), https://newslab.withgoogle.com/assets/docs/data-journalism-in-2017.pdf [https://perma.cc/P69K-8UMM].

247. See Mirko Lorenz, Why Journalists Should Use Data, *in* 1 The Data Journalism Handbook 3, 3 (Liliana Bounegru, Lucy Chambers & Jonathan Gray eds., 2012), https://datajournalismhandbook.org/uploads/first_book/DataJournalismHandbook-2012.pdf [https://perma.cc/C2KG-LFFZ] (describing how data journalism connects discrete points of information that "are often not relevant in a single instance, but massively important when viewed from the right angle").

248. See supra notes 1–16 and accompanying text; see also Baranetsky, supra note 12 ("[C]ompanies rarely make this data available to universities, researchers, or think tanks, limiting the ability of outside institutions to hold them accountable for misuse or educate users about the way they may be receiving influenced information.").

249. See Dep't of Justice, supra note 118, at 12–57 (describing the various methods of prosecuting individuals for alleged misuse of data under the CFAA).

250. Baranetsky, supra note 12.

criminal penalties has deterred academic research and caused journalists to withhold stories.[251]

There is a demonstrated need for CFAA reform in order to allow stories and research that serve the public interest to proceed without opening up members of the public to harmful scraping activity.[252] This Part argues that a resolution to the inherent tensions in assessing CFAA reform from a First Amendment perspective is best approached by drawing a fine line around the scope of scraping activity intended to be protected. First, section III.A proposes a legislative solution for CFAA reform, including adding a safe harbor provision to the statute and creating a regulatory body to oversee digital research. Next, section III.B advocates for narrow statutory interpretation of the CFAA in order to minimize imposition of liability on journalists and researchers. Each section also addresses potential criticism of such reforms.

A. Legislative Reform: A Safe Harbor Provision and Regulatory Body

Because there is an internal inconsistency in assessing the CFAA solely from a First Amendment perspective, a legislative solution will need to comport with both First Amendment interests of accountability and political self-governance, on the one hand, and privacy, on the other. One previous legislative reform effort was Aaron's Law, a bill aimed at narrowing the scope of the CFAA so that mere violations of websites' terms of

---

251. See, e.g., Camille Fassett, Facebook's Terms of Service Obstruct Important Journalistic Research, Freedom of the Press Found. (Aug. 7, 2018), https://freedom.press/news/facebooks-terms-service-obstruct-important-journalistic-research/ [https://perma.cc/3HTX-GLHU] (describing how PBS journalist Cameron Hickey's proposed project identifying unknown sources of misinformation on Facebook was shut down due to fear of legal liability for scraping); Kashmir Hill & Surya Mattu, Facebook Wanted Us to Kill This Investigative Tool, Gizmodo (Aug. 7, 2018), https://gizmodo.com/facebook-wanted-us-to-kill-this-investigative-tool-1826620111 [https://perma.cc/K5PQ-TMPQ] (describing Facebook's attempt to shut down two Gizmodo journalists' scraping tool that kept track of Facebook's "People You May Know" suggestions to users); see also Letter from Jameel Jaffer, Exec. Dir., Knight First Amendment Inst. at Columbia Univ., to Mark Zuckerberg, CEO, Facebook (Aug. 6, 2018), https://knightcolumbia.org/sites/default/files/content/Facebook_Letter.pdf [https://perma.cc/3LJQ-6KWU] ("We have spoken to a number of journalists and researchers who have modified their investigations to avoid violating Facebook's terms of service, even though doing so made their work less valuable to the public. In some cases, the fear of liability led them to abandon projects altogether.").

252. See, e.g., Baranetsky, supra note 12 ("[I]t is increasingly likely that companies will . . . make various information even more difficult to access[,] . . . triggering an ever greater need for reporters and academics to continue their work advocating for more transparency and access . . . as algorithms and data sets become distinctly determinative factors in our lives."). But see Letter from Jameel Jaffer to Mark Zuckerberg, supra note 251, at 3 ("We understand that, in the wake of revelations concerning the Cambridge Analytica scandal, Facebook is facing new pressure to protect the data that users entrust to it. This pressure is warranted and indeed overdue.").

service would not result in federal charges.[253] The bill proposed removing the phrase "exceeds authorized access" and replacing it with "access without authorization" in order to clarify that a person cannot be punished merely for violating a website's terms of service.[254] Under Aaron's Law, redundant provisions that allow a person to receive multiple punishments for the same action would be eliminated from the CFAA.[255] However, the bill stalled and has languished in Congress for several years.[256] Additionally, while the bill would limit the potential for journalists and researchers to be prosecuted for scraping, it would not eliminate the possibility of civil and criminal repercussions entirely. This Note proposes legislative reforms specifically aimed at allowing journalists and researchers to engage in scraping activity without repercussions.

The first proposed reform is the creation of a safe harbor provision for scraping in the public interest conducted by academics, researchers, and journalists—that is, scraping that falls in the taxonomy classification outlined in section I.A.2.b. The safe harbor provision would permit journalists and researchers to conduct investigations in the public interest and utilize scraping techniques that may otherwise violate a website's terms of service and thus subject the researcher to liability under the CFAA. However, scrapers not serving the public interest[257] would not be entitled to such a safe harbor in an effort to minimize any resulting privacy harms.

The Knight First Amendment Institute (Knight Institute) recently proposed a similar safe harbor provision on which this Note's CFAA-reform proposal is based.[258] In the summer of 2018, the Knight Institute sent a letter to Facebook CEO Mark Zuckerberg proposing an amendment to Facebook's terms of service.[259] Under Facebook's current user agreement, researchers are prevented from utilizing scraping techniques such as "automated collection of public information and the creation of temporary research accounts."[260] The Knight Institute's proposed amendment

253. See Zoe Lofgren & Ron Wyden, Introducing Aaron's Law, A Desperately Needed Reform of the Computer Fraud and Abuse Act, WIRED (June 20, 2013), https://www.wired.com/2013/06/aarons-law-is-finally-here/ (on file with the *Columbia Law Review*).

254. See Aaron's Law Act of 2015, S. 1030, 114th Cong. § 2(b)(1) (as read and referred to the S. Comm. on the Judiciary, Apr. 21, 2015); Aaron's Law Act of 2013, H.R. 2454, 113th Cong. § 2(b)(1) (as referred to the H. Subcomm. on Crime, Terrorism, Homeland Sec., and Investigations, July 15, 2013).

255. See S. 1030 § 3; H.R. 2454 § 3.

256. See Thomas Brewster, Aaron's Law Is Doomed Leaving US Hacking Law 'Broken,' Forbes (Aug. 6, 2014), https://www.forbes.com/sites/thomasbrewster/2014/08/06/aarons-law-is-doomed-leaving-us-hacking-law-broken/ [https://perma.cc/YF8H-35T4].

257. See supra section I.A.2.a (describing commercial and other uses of scraping).

258. See Letter from Jameel Jaffer to Mark Zuckerberg, supra note 251, app.

259. Id. at 1–5, app.

260. Id. at 2; see also Terms of Service, Facebook, https://www.facebook.com/legal/terms/update [https://perma.cc/AWC7-K728] (last updated July 31, 2019) ("You may not

would create a safe harbor for journalists and academics who wish to investigate the social media site.[261]

The Knight Institute's letter highlighted four key aspects of reforming Facebook's terms. First, to qualify for safe harbor protection, the project must be "to inform the general public about matters of public concern."[262] The project specifically cannot "facilitate commercial data aggregation [or] targeted advertising."[263] Second, the project must take "reasonable measures" to protect Facebook users' privacy.[264] This means that data obtained from scraping must be stored securely and cannot be used for any purpose other than informing the general public.[265] As such, it must not be sold, licensed, or transferred to another party.[266] Scrapers also cannot disclose "any information that would readily identify a user without the user's consent" unless there is a compelling public interest in such disclosure.[267] Third, the project also cannot be unreasonably misleading to Facebook users.[268] Fourth, the project cannot be disruptive or burdensome to Facebook's platform either in functionality and appearance.[269] The Knight Institute's suggested safe harbor provision stated that researchers do not violate the platform's terms of service by engaging in activity that meets these four requirements.[270]

The Knight Institute's proposal to Facebook is an effective one: It walks the fine line between ensuring access for scraping that would advance First Amendment values of accountability and self-governance while also preventing a data scraping free-for-all that would infringe upon First Amendment values of privacy and autonomy. However, the Knight Institute's safe harbor focuses only on Facebook.[271] Although Facebook is one of the largest social media networking sites—and one that has been in the spotlight following the 2016 presidential election—the proposal does

---

access or collect data from our Products using automated means (without our prior permission) or attempt to access data you do not have permission to access.").

261. Letter from Jameel Jaffer to Mark Zuckerberg, supra note 251, app.

262. Id. at app. § 1(1).

263. Id. at 4.

264. Id.

265. Id.

266. Id.

267. Id. at app. § 1(2).

268. Id. at app. § 1(3) ("A temporary research account should generally be identified as such in the account's public profile.").

269. See id. at app. § 1(4).

270. See id. at app. § 1.

271. See id.

not account for any other websites.[272] By incorporating a similar safe harbor provision directly into the CFAA, journalists and researchers would not have to rely on the individual buy-in from each website to adopt such terms.[273]

While the safe harbor provision would be a big step in the right direction for long-overdue CFAA reform, simply adding such a provision may not be enough to satiate reformers. One potential critique is that a safe harbor provision gives too much control to websites to decide who deserves protection as a journalist or researcher.[274] Indeed, Facebook faced similar scrutiny over its plans to rank "trusted" news outlets.[275] One way of coping with such ambiguities would be to create a regulatory body that would govern the uses of scraping and classify requests under the safe harbor provision. This body would be akin to an institutional review board at a university, which reviews methods proposed for research to ensure that they are ethical and approves whether research can proceed.[276]

A second potential shortcoming concerns what happens when platforms erect technical barriers to access. Courts have suggested that in some situations, code-based barriers that restrict access to information do not violate the First Amendment. For example, in *Sandvig v. Sessions*, the D.C. District Court suggested that if a company naturally strengthens its code such that scraping is not technologically feasible on its platform, such

---

272. See id.

273. Facebook has publicly addressed its reservations about the safe harbor proposal, noting that Facebook has "strict limits" for third parties using personal information. See Christine Schmidt, If Facebook Makes a Safe Harbor for Journalists and Researchers, Would It Help?, Nieman Lab (Aug. 7, 2018), http://www.niemanlab.org/2018/08/if-facebook-makes-a-safe-harbor-for-journalists-and-researchers-would-it-help/ [https://perma.cc/L8C5-N8RM]. To date, Facebook has not amended its terms of service to incorporate the proposed safe harbor provision. See Facebook, Terms of Service, supra note 260. This demonstrates the difficulty of getting corporations to voluntarily sign on and suggests that such reforms would be more effective if mandated by the government.

274. See Mathew Ingram, Should Journalists and Researchers Get a Special Exemption from Facebook's Rules?, Colum. Journalism Rev. (Aug. 7, 2018), https://www.cjr.org/the_new_gatekeepers/facebook-journalists-researchers-rules.php [https://perma.cc/CPQ6-6N5B] ("[W]ho gets to decide who is deserving of protection? Having Facebook choose which researchers qualify might not raise too many red flags, but giving a private corporation the ability to say who is or isn't a journalist would be hugely controversial . . . ."). The Knight Institute has responded, noting that the proposal doesn't ask Facebook to decide who is and is not a journalist; instead, the proposal asks Facebook to make a decision based on the purpose of the project. Id.

275. See Mathew Ingram, Campbell Brown on Facebook's Plans to Decide What News Is Trustworthy, Colum. Journalism Rev. (May 3, 2018), https://www.cjr.org/q_and_a/campbell-brown-facebook-news.php [https://perma.cc/UL4W-E2Z4] (describing Facebook's uncertain approach to defining a "broadly trusted" news organization).

276. See, e.g., Food & Drug Admin., Institutional Review Boards Frequently Asked Questions: Guidance for Institutional Review Boards and Clinical Investigators (Jan. 1998), https://www.fda.gov/RegulatoryInformation/Guidances/ucm126420.htm [https://perma.cc/9D26-GDF6].

technological prohibitions of scraping may not violate the First Amendment.[277] However, recent jurisprudence suggests that mere technological barriers that simply limit the feasibility of scraping without creating a code-based delineation of public spaces from private spaces online do not withstand legal challenges.[278] For example, in *hiQ Labs v. LinkedIn*, the Ninth Circuit upheld a district court order mandating that LinkedIn remove technical barriers to access to public profiles because LinkedIn's barriers didn't operate on a private space.[279] The steps LinkedIn took to restrict access to publicly available user profiles simply detected "suspicious activity" and restricted automated scraping.[280] This type of access restriction is distinct from access that requires affirmative authentication requirements, such as a password gate.[281]

The Ninth Circuit *hiQ* opinion is instructive on how courts should view attempts to circumvent technological barriers to access: Courts should examine whether the conduct is analogous to breaking and entering because the scraping attempts to circumvent barriers to virtual private space.[282] For example, an attempt to circumvent authentication requirements—such as a password gate—is clearly beyond the scope of protected activity, regardless of whether the person is scraping to serve the public interest or for commercial uses because the norm of access shifts from open to closed.[283] However, scraping techniques that evolve to work around existing code when there is no barrier or affirmative authentication requirement delineating public from private spaces online should not be deemed prohibited conduct.

---

277. 315 F. Supp. 3d 1, 13 (D.D.C. 2018) ("[O]nly code-based restrictions, which 'carve[] out a virtual private space within the website or service that requires proper authentication to gain access,' remove those protected portions of a site from the public forum." (quoting Orin S. Kerr, Norms of Computer Trespass, 116 Colum. L. Rev. 1143, 1171 (2016))).

278. See hiQ Labs, Inc. v. LinkedIn Corp., No. 17-16783, 2019 WL 4251889, at *12 (9th Cir. Sept. 9, 2019).

279. See id. at *4, *15.

280. See id. at *2 ("For example, LinkedIn's Quicksand system detects non-human activity indicative of scraping; its Sentinel system throttles (slows or limits) or even blocks activity from suspicious IP addresses; and its Org Block system generates a list of known 'bad' IP addresses serving as large-scale scrapers." (footnote omitted)).

281. See id. at *11–12 (describing how the CFAA originally guarded against hacking when "affirmative authorization" was required for access and how scraping publicly available LinkedIn profiles is not sufficiently analogous to breaking and entering).

282. See id. at *12 ("The legislative history of section 1030 thus makes clear that the prohibition on unauthorized access is properly understood to apply only to private information—information delineated as private through use of a permission requirement of some sort.").

283. See id. Professor Orin S. Kerr has compared online technological barriers to access to a store that is "open to the public in the front but for employees only in the back." Kerr, supra note 277, at 1171. An authentication requirement "imposes a barrier that overrides the Web default of open access." Id.

While a safe harbor provision may not provide access for journalists and researchers in all situations when platforms erect technological barriers to access, it is worth noting that scraping techniques are the product of diligent developments and innovation by journalists and technologists.[284] As long as websites develop their code, it is likely that journalists and researchers will similarly evolve to keep up with current trends in a way that does not constitute breaking and entering. This Note does not purport to argue that journalists and researchers should have special access to information otherwise in a private forum; such a FOIA-for-private-corporations project is beyond this Note's scope. Instead, this Note suggests that when the information is obtainable, journalists and researchers should not be subject to legal liability for using digital tools to obtain such information.

B. Narrow Statutory Interpretation

Although not as comprehensive as legislative reform, adopting the narrow statutory interpretation of the CFAA access provision followed by the Second, Fourth, and Ninth Circuits[285] would also minimize the legal risk for journalists and academics hoping to scrape for their research projects. As described in section I.B above, under the narrow interpretation of the access provision, access is only penalized if it amounts to "breaking and entering" a computer; a mere terms of service violation without more would be insufficient to subject researchers to liability under the CFAA.[286] This approach is illustrated in *Sandvig v. Sessions*.[287] However, although the *Sandvig* court declined to carve out an exception from the CFAA for harmless terms of service violations,[288] the court did adopt a narrow approach and noted the risks of enforcing the CFAA against journalists and researchers.[289] Under the narrow interpretation approach as

284. See Eric Johnson, It May Be 'Data Journalism,' but Julia Angwin's New Site the Markup Is Nothing Like FiveThirtyEight, Recode (Sept. 27, 2018), https://www.recode.net/2018/9/27/17908798/julia-angwin-markup-jeff-larson-craig-newmark-data-investigative-journalism-peter-kafka-podcast [https://perma.cc/8MDN-72GK] (describing how journalists have developed and built the tools for scraping investigations).

285. See notes 132–135 and accompanying text (describing the narrow interpretation approach followed by some circuits).

286. See note 136 and accompanying text.

287. See 315 F. Supp. 3d. 1, 22–27 (D.D.C. 2018).

288. See id. at 23 ("It is one thing to carve out such violations by determining that the statute is unconstitutional as applied, but the text of the statute itself—'exceeds authorized access'—and its statutory definition do not appear to allow for such a surgical slicing off of conduct.").

289. See id. at 23–26 ("By incorporating [terms of service] that purport to prohibit the purposes for which one accesses a website or the uses to which one can put information obtained there, the CFAA threatens . . . a great deal of expressive activity, even on publicly accessible websites—which brings the First Amendment into play.").

articulated in *Sandvig*, the court appeared to retain flexibility to balance the contemplated harm with the benefit conferred.

However, there are some apparent shortcomings of relying on a narrow statutory interpretation to ensure access for journalists and researchers. Most notably, the *Sandvig* court explicitly stated that the central inquiry in interpreting the CFAA is what information the parties plan to access via scraping, "not on why they wish to access it, the manner in which they use their authorization to access it, or what they hope to do with it."[290] This caveat implies that the scraper must only access information that occurs on portions of websites that any visitor can view.[291] Such a limitation would prohibit the research tactic of creating sock-puppet accounts (that is, fictitious user accounts).[292] Still, the court appeared to offer a way out of this dilemma by noting that creating fake profiles may also be First Amendment–protected activity, even if it technically falls within the ambit of the CFAA.[293] Therefore, it's possible that adopting the narrow interpretation approach could provide meaningful protection to journalists and researchers on constitutional avoidance grounds. Additionally, notwithstanding the court's comments about intended use, some following the case have argued that the court's opinion was influenced by the academic purpose of the scraping and that it would have reached a different outcome had it been a commercial case.[294] Therefore, there may still be some leeway for leniency in the use of fictitious accounts based on intended purpose of use.

---

290. Id. at 26.

291. See id. at 26–27 ("Scraping or otherwise recording data from a site that is *accessible to the public* is merely a particular use of information that plaintiffs are entitled to see." (emphasis added)).

292. See id. at 27 ("Out of plaintiffs' proposed activities, then, only Mislove and Wilson's plan to create fictitious user accounts on employment sites would violate the CFAA."). For further discussion of this research technique, see supra notes 49–53 and accompanying text.

293. See *Sandvig*, 315 F. Supp. 3d at 30 ("At this stage, 'absent any evidence that the speech [would be] used to gain a material advantage,' . . . plaintiffs' false speech on public websites retains First Amendment protection . . . and rendering it criminal does not appear to advance the government's proffered interests." (quoting United States v. Alvarez, 567 U.S. 709, 723 (2012))).

294. See, e.g., Jeffrey Neuburger, Researchers May Challenge the Constitutionality of the CFAA "Access" Provision as Applied to Web Scraping, Proskauer (Apr. 27, 2018), https://newmedialaw.proskauer.com/2018/04/27/researchers-may-challenge-the-constitutionality-of-the-cfaa-access-provision-as-applied-to-web-scraping/ [https://perma.cc/7BFP-ZYPC] ("Presumably, the court would have reached a different result if the plaintiffs were not professors or big data researchers test auditing websites as opposed to data scrapers seeking commercial gain or otherwise engaging in what a website owner might deem 'free riding.'").

CONCLUSION

As the internet affects the day-to-day life of nearly every member of society and platforms like Facebook influence public discourse in increasingly important ways, journalists and researchers play a crucial role in reporting on and illuminating this influence on the public. Digital research techniques like scraping can serve First Amendment values of self-governance and democracy by contributing to an informed public and exposing the role that big data and social media platforms play in society.[295] Yet journalists and researchers are subject to criminal and civil liability under the CFAA for scraping; protected First Amendment activity is thus chilled by this statute. Further complicating the equation is the idea that such scraping techniques can simultaneously implicate First Amendment values of intellectual privacy and autonomy by allowing players to harvest personal data for commercial and potentially manipulative uses. Thus, this Note suggests that the CFAA is in dire need of reform to permit scraping by journalists and researchers in their public-interest-oriented research projects while protecting against privacy abuses and security risks.

The CFAA is a clunky statute that has been the target of many reform proposals in recent years, to little avail.[296] However, such proposed reforms have not contemplated a solution that would differentiate based on different forms of data scraping. Allowing researchers and reporters to investigate matters of public interest while still protecting the valid privacy concerns raised by such practices would vindicate First Amendment values.

---

295. See supra sections I.A.1–.2.

296. See supra notes 253–256 and accompanying text; see also Computer Fraud and Abuse Act Reform, Elec. Frontier Found., https://www.eff.org/issues/cfaa [https://perma.cc/L5DX-5SZY] (last visited Sept. 15, 2019) (proposing reforms of the CFAA).