WASHINGTON JOURNAL OF LAW, TECHNOLOGY & ARTS
VOLUME 13, ISSUE 3 SPRING 2018

# ROBOTS WELCOME? ETHICAL AND LEGAL CONSIDERATIONS FOR WEB CRAWLING AND SCRAPING

*Zachary Gold and Mark Latonero*[*]


Cite as: 13 Wash. J.L. Tech. & Arts 275 (2018)

http://digital.law.washington.edu/dspace-law/handle/1773.1/1817

## ABSTRACT

*Web crawlers are widely used software programs designed to automatically search the online universe to find and collect information. The data that crawlers provide help make sense of the vast and often chaotic nature of the Web. Crawlers find websites and content that power search engines and online marketplaces. As people and organizations put an ever-increasing amount of information online, tech companies and researchers deploy more advanced algorithms that feed on that data. Even governments and law enforcement now use crawlers to carry out their missions. Despite the ubiquity of crawlers, their use is ambiguously regulated largely by online social norms whereby webpage headers signal whether automated "robots" are welcome to crawl their sites. As courts take on the issues raised by web crawlers, user privacy hangs in the balance. In August 2017, the Northern District of California granted a preliminary injunction in such a case, deciding that*

---

[*] Zachary Gold works for the New York City Mayor's Office of the Chief Technology Officer. He previously researched privacy and human rights issues at the Data & Society Research Institute, a non-profit independent research organization in New York that studies the social and cultural issues arising from data-centric and automated technologies. He obtained his J.D. from the American University Washington College of Law and his M.A. in Public Policy from SUNY Stony Brook. Dr. Mark Latonero (Ph.D. University of Southern California) leads the Human Rights Program at the Data & Society Research Institute. He is a fellow at UC Berkeley School of Law's Human Rights Center as well as the USC Annenberg School and Leiden University in the Netherlands. The views expressed in this article are those of the authors and do not reflect the views of the City of New York.

*LinkedIn's website must be open to such crawlers. In March 2018, the District Court for the District of Columbia granted standing for an as-applied challenge to the Computer Fraud and Abuse Act to a group of academic researchers and a news organization. The Court allowed them to proceed with a case in which they now allege the law's making a violation of website Terms of Service a crime effectively prohibits web crawling and infringes on their First Amendment Rights. In addition, news media is inundated with stories like Cambridge Analytica wherein web crawlers were used to scrape data from millions of Facebook accounts for political purposes.*

*This paper discusses the history of web crawlers in courts as well as the uses of such programs by a wide array of actors. It addresses ethical and legal issues surrounding the crawling and scraping of data posted online for uses not intended by the original poster or by the website on which the information is hosted. The article further suggests that stronger rules are necessary to protect the users' initial expectations about how their data would be used, as well as their privacy.*

TABLE OF CONTENTS

## INTRODUCTION

Scientists, researchers, private industry, and government are tuning in to the changes in information-gathering and analysis brought about by big data. Where relatively small data projects—public opinion surveys, questionnaires, or other similar projects—were once used to provide answers to scientific, business, and civic questions, we can now turn to the much larger store of information on the Internet to try to find better or faster answers to those same questions. Using algorithms and artificial intelligence, we can increase efficiency, augment labor, and complete tasks that are too massive, complicated, or otherwise difficult for humans to realistically complete.

Private companies like Google, Microsoft, and others have for decades provided answers—or, more commonly, provided a list of locations where one might find an answer. They use web crawlers to search and index the web to provide reliable, relevant web pages in response to search queries.[1] Further, these algorithms index a relatively small portion of the worldwide web,[2] and much less of the broader internet. Not only do these crawlers search a limited number of websites, they also save little information from them. Search engines tend to care only about which websites link to which other websites, maintaining headlines and snippets of text to display to users, or saving thumbnails of images for the same reason. Much of

---

[1] *See e.g.*, *How Google Search Works*, GOOGLE, https://support.google.com/webmasters/answer/70897?hl=en (last visited May 1, 2018).

[2] *See* Andy Beckett, *The Dark Side of the Internet*, THE GUARDIAN (Nov. 25, 2009), https://www.theguardian.com/technology/2009/nov/26/dark-side-internet-freenet.

the data stored on the web is ignored by crawlers entirely, and not scraped for indexing and searching.[3] But this data is the raw material for big data analytics, machine learning algorithms, and similar tools that attempt to analyze, inform, and predict.

While web crawlers are mostly used to collect the relatively limited information necessary to power search engines, they can be used to search, index, and later analyze vast amounts of information on the internet. Increased storage capabilities and computing power are making such usage more practical. Governments can use web crawlers to find criminals operating online. Researchers can use them to identify social trends or political opinions. Private companies may try to glean information about their customers and their preferences from data scrapped from forums, blogs, social media websites, or elsewhere.

These basic functions, long used for well understood purposes, will soon be—or are already being—used to provide the raw data for analyses that many may consider uncomfortable, unethical, or even illegal. They can provide the images necessary to feed a facial recognition system, the content needed to search for violent extremists, or to jump-start a business using data someone else already collected.

This raises a number of questions about the use of such software and the status of the websites they crawl. For this reason, a number of institutions have sought to address this issue. The American Association for Public Opinion Research published its own report identifying data ownership, data stewardship, data collection authority, privacy and reidentification, and data protection as policy challenges to be addressed.[4] The White House, under President Obama, also released a report on big data discussing government uses and providing a background on U.S. privacy law, ranging from

---

[3] *See* J.J. Rosen, *The Internet You Can't Google*, TENNESSEAN, https://www.tennessean.com/story/money/tech/2014/05/02/jj-rosen-popular-search-engines-skim-surface/8636081/ (last updated May 3, 2014) (reporting that Google indexes "only an estimated 4 percent of the information that exists on the Internet.").

[4] *See* Lilli Japec et al., *AAPOR Report: Big Data*, AM. ASS'N FOR PUB. OPINION RES. (Feb. 12, 2015), https://www.aapor.org/Education-Resources/Reports/Big-Data.aspx#3.2%20Paradigm%20Shift.

Samuel Warren and Louis Brandeis' *The Right to Privacy*, to the Fair Information Practice Principles and the Consumer Privacy Bill of Rights.[5] The report, among other things, discussed big data's effect on citizenship, discrimination, and privacy, and made a number of general recommendations, including a national data breach standard, developing technical expertise to stop discrimination, and amending the Electronic Communications Privacy Act (ECPA).[6]

Prior discussions have failed to provide implementable technology or policy solutions, leaving many questions unanswered. In the context of government use, can crawling and scraping ever constitute a search or seizure that would be governed by the Fourth Amendment? More broadly, as applied to the private sector and researchers, do internet users have a privacy interest in what they post online? How and when does such an interest operate? What kind of policies should crawlers obey to protect those searched? Do current federal laws apply to these activities, and do they have the necessary force to meaningfully protect internet users' data from being made part of a database that will be used for purposes users did not or could not foresee?

Technology often advances ahead of law and policy. Web crawlers are currently governed almost entirely by social norms and politeness, and neither Congress, the executive branch, nor the courts have promulgated laws or guidelines specifically governing their use as tools of surveillance. Without any such rules, there is a near certainty that someone's privacy has already been, or will soon be violated, their statements connected to their true identity, online posts used against them in court, or some unforeseen harm caused. This article will discuss the problems raised by big data and web crawling from an ethical and legal standpoint. The question of how to regulate crawling and scraping data with bots by government, the private sector, researchers, and individuals will be examined with the goal of identifying issues and highlighting specific dilemmas for policymakers to address before widespread surveillance using web

---

[5] EXECUTIVE OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES, PRESERVING VALUES, (May 2014), https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

[6] *Id.* at 60.

crawlers can cause undue harm.

## I. WEB CRAWLERS

Web Crawlers, also called bots, spiders, and crawlers are in common use on the web. Perhaps of most familiarity to the average internet user, their work product is on display whenever one uses a search engine like Google. Search engines employ crawlers to systematically scan, analyze, and save information about websites to index those sites for searching, determine their importance to a particular search, and find connections between websites.[7]

Web crawlers visit websites at the direction of their operators, but often with little everyday input from them. Operators may choose all the web pages that a crawler will visit, but more often they are driven by algorithms making that determination. For example, Googlebot, the web crawler Google uses to inform its search engine, uses an algorithm to determine what to crawl based on data from previous crawls.[8] These crawlers may visit a given web page a number of times a day to ensure data is collected in a timely fashion.[9] Often, there is a way for website operators to submit their URLs manually to request that a bot crawl their websites.[10] Nevertheless, crawls are often automatic and informed by the sample of the web searched, necessitating that some websites will be left out, and leading to some amount of bias in the results of the crawl. Web crawlers can provide information in real time.[11]

Because crawlers are so active and bandwidth is limited, crawlers generally operate politely, in accordance with social

---

[7] *See* Javed Mostafa, *How do Internet Search Engines Work?*, SCIENTIFIC AMERICAN (Oct. 14, 2002) https://www.scientificamerican.com/article/how-do-internet-search-en/.

[8] *See Googlebot*, GOOGLE, https://support.google.com/webmasters/answer/182072?hl=en (last visited May 1, 2018).

[9] *See What Are Crawlers? How Do They Work?*, SEO MARKETING WORLD, http://www.seomarketingworld.com/seo-faq/crawlers.php (last visited May 1, 2018).

[10] *See How Does a Robot Decide Where to Visit?*, ROBOTSTXT.ORG, http://www.robotstxt.org/faq/visit.html (last visited May 1, 2018).

[11] *See* David Harry, *Crawling and the Real Time Web*, SEJ (Apr. 29, 2010), https://www.searchenginejournal.com/crawling-and-the-real-time-web/20510/.

norms—the desires of website operators are stored in the code of their websites. Crawlers, poorly designed or left to run freely, can use significant network resources or even crash servers.[12] For this reason, a protocol exists to temper the crawls performed by these bots. Website administrators use the Robots Exclusion Protocol, or "robots.txt", to ask crawlers not to search particular pages of their website, or to leave it un-crawled entirely.[13] This file can be targeted at specific bots (for example, telling only Googlebot not to index a page), or at all bots.[14] Some robots will also respect requests to time delays between crawls to conserve network resources.[15] However, robots.txt can be ignored; those employing crawlers are not bound by any law, contract, or technical need to obey a robots.txt file.[16] Only politeness and social pressure provide enforcement power. There are other methods of keeping crawlers out, such as requiring users to log in, or fill in a captcha, but those too can be sidestepped by a bot's programmers.[17]

As technology advances, web crawlers are able to scrape more data from websites. Where it may not have been possible to save all the text or images from a website in the past, as the cost of storage has gone down, the operators of a web crawler can now scrape and store far more information, including comments and the identities of those who posted them, advertisements, and pictures. Advancements in facial recognition technology allow people in images to be identified, and disparate online identities can be

---

[12] *See Aren't Robots Bad for the Web?*, ROBOTSTXT.ORG, http://www.robotstxt.org/faq/bad.html (last visited May 1, 2018) ("Certain robot implementations can (and have in the past) overloaded networks and servers. This happens especially with people who are just starting to write a robot; these days there is sufficient information on robots to prevent some of these mistakes.").

[13] *See About /robots.txt*, ROBOTSTXT.ORG, http://www.robotstxt.org/robotstxt.html (last visited May 1, 2018) (explaining how to use robots.txt to allow robots complete access, exclude robots entirely, exclude or allow particular robots, or how to disallow crawling of particular pages).

[14] *Id.*

[15] *See Robots.txt Tutorial*, SEOBOOK, http://tools.seobook.com/robots-txt/ (last visited May 1, 2018).

[16] *Can a /robots.txt Be Used in a Court of Law?*, ROBOTSTXT.ORG, http://www.robotstxt.org/faq/legal.html (last visited May 1, 2018).

[17] *See, e.g.*, Tim Anderson, *How Captcha Was Foiled: Are You a Man or a Mouse?*, THE GUARDIAN (Aug. 27, 2008), http://www.theguardian.com/technology/2008/aug/28/internet.captcha.

connected to a real person.

## II. PRIVACY CONCERNS

Web crawlers provide the ability for any sufficiently sophisticated and funded operator to maintain a fairly ubiquitous surveillance regime over a larger number of internet domains. This has serious implications for the privacy of internet users. Web crawlers can be used for widespread tracking of internet users without their knowledge or consent. When paired with other technologies, these crawlers can successfully deanonymize people who post online under pseudonyms, or even identify people who have merely had pictures of them posted by others.

Web crawlers can be used to easily acquire large amounts of information, including who posts on which websites, who they interact with, and what they post. This may reveal political, religious, and other views of users, along with significant personal information. Some government agencies already use various methods to track protests and protesters,[18] and eight out of ten law enforcement professionals use social media as a tool in their investigations.[19] Web crawlers enable government agents to quickly collect data from web forums, personal blogs, social networking sites like Twitter, Facebook, and Tumblr, or bulletin boards like Craigslist. Web crawlers also allow government agents to collect data from protest groups' websites to determine the number of

---

[18] *See, e.g.*, George Joseph, *Exclusive: Feds Regularly Monitored Black Lives Matter Since Ferguson*, THE INTERCEPT (July 24, 2015), https://firstlook.org/theintercept/2015/07/24/documents-show-department-homeland-security-monitoring-black-lives-matter-since-ferguson (explaining that the Department of Homeland Security collected information, "including location data . . . from public social media accounts, including on Facebook, Twitter, and Vine, even for events expected to be peaceful. . . . They also show the department watching over gatherings that seem benign and even mundane. . . . [A] DHS-funded agency planned to monitor a funk music parade and a walk to end breast cancer in the nation's capital.").

[19] *See Social Media Use in Law Enforcement: Crime Prevention and Investigative Activities Continue to Drive Usage*, LEXISNEXIS, at 2 (Nov. 2014), https://risk.lexisnexis.com/-/media/files/government/white-paper/2014-social-media-use-in-law-enforcement-pdf.pdf [hereinafter LEXISNEXIS].

protestors, identify the protestors, and discover their motivations.[20] These activities have important constitutional implications as they could chill protected speech, infringe on protester's freedom of association, or violate a person's Fourth Amendment right to protection against unreasonable searches. Corporations and researchers are also using crawlers to scrape internet data to inform their business practices and research.[21] While these corporate practices do not implicate the same constitutional rights as government use of crawlers, they do have significant bearing on the privacy rights of internet users whose data is collected. Not only might the initial collection by corporations or researchers violate the privacy of internet users, but poor security practices could result in data breaches putting personal data in the hands of people with malicious motives.

This collection of information can be done without the knowledge or consent of those posting. Users post online with certain expectations about how their posts will be used, and while they may use websites that include privacy controls or have terms of service (ToS) forbidding crawling, these may be circumvented. Privacy controls are often too confusing for users to employ effectively,[22] and in any case do not control what others post. And, as discussed above, very little controls the ability of web crawlers to scrape data from a web page.[23] This means that government

---

[20] *See* Richard Esposito et al., *Showden Docs Reveal British Spies Snooped on YouTube and Facebook*, NBC NEWS (Jan. 27, 2014), http://investigations.nbcnews.com/_news/2014/01/27/22469304-snowden-docs-reveal-british-spies-snooped-on-youtube-and-facebook. In 2012, the British Government Communications Headquarters demonstrated the ability to monitor YouTube, Facebook, and Twitter in real time; this sort of information apparently has value to governments interested in monitoring online activity.

[21] *See* discussion *infra* Part V.

[22] *See* Josh Constine, *Facebook Admits Users Are Confused About Privacy, Will Show More On-Screen Explanations*, TECHCRUNCH (Apr. 8, 2014), http://techcrunch.com/2014/04/08/facebook-privacy-settings/ ("Facebook's privacy team manager Mike Nowak admitted that people think Facebook changes its privacy controls too often or that the company has failed to make privacy easy to understand.").

[23] *See, e.g.*, *How Do I Prevent Robots Scanning My Site?*, ROBOTS.TXT, http://www.robotstxt.org/faq/prevent.html (last visited May 1, 2018) (providing advice on how to prevent scraping by crawlers, but noting "this only helps with

agencies, corporations, or others can easily navigate around users' expectations, collecting whatever data they want without the subjects of the surveillance ever learning of the collection, much less having a chance to consent.

This sort of tracking, scraping, and storage of information allows governments to engage in further invasions of privacy beyond merely collecting information on individuals as they interact both online and offline. Such practices have serious implications for unmasking real identities online.

Facial recognition technology can, to varying degrees, accurately identify a person in a picture.[24] This allows a government agency, or others, to scrape images from websites to identify the people in the photos, creating a database of users, their acquaintances, and friends. Because metadata is often uploaded with such photos, the times and locations of users' meetings may also be collected. To an increasing extent, clear images of peoples' faces are not necessary as computers are being trained to identify people based on factors like hair style, clothing, body shape, and pose.[25] Users cannot avoid this sort of surveillance by refraining from taking pictures of themselves, or by asking their friends not to post photos or tag them. It is possible that images posted by strangers may lead to ones' identification in the background of a picture with an entirely different subject.

Such crawling and scraping can also be used to unmask aliases. Crawlers may scrape information like physical addresses, email addresses, phone numbers, or linked accounts that can be used to link aliases to each other, or to link an alias to a real-world identity, stymying attempts to speak anonymously. While this is certainly

---

well-behaved robots.").

[24] *See* Russell Brandom, *Why Facebook is Beating the FBI at Facial Recognition*, THE VERGE (July 7, 2014), http://www.theverge.com/2014/7/7/5878069/why-facebook-is-beating-the-fbi-at-facial-recognition; *see also* James Geddes, *Windows 10 Hello Facial Recognition Feature Can Distinguish Between Identical Twins*, TECH TIMES (Aug. 25, 2015), http://www.techtimes.com/articles/79108/20150825/windows-10-hello-facial-recognition-feature-can-distinguish-between-identical-twins.htm (describing a small test undertaken by a journalist).

[25] Aviva Rutkin, *Facebook can Recognise You in Photos Even if You're Not Looking*, NEW SCIENTIST (June 22, 2015), https://www.newscientist.com/article/dn27761-facebook-can-recognise-you-in-photos-even-if-youre-not-looking#.VYjUthNVhBd.

possible without crawlers, crawlers' ability to search constantly and systematically increases the chances that a user's mistake or private material will be found and taken advantage of. Further, this can be done on a large scale, leading to the potential unmasking of a great number of aliases.

Crawlers are accessible to nearly anyone with a bit of technical expertise and access to the necessary computing resources to complete their task. While government crawling and scraping has implications for the privacy as well as the First and Fourth Amendment rights of U.S. citizens, application of these tools by private entities is not without risks.

## III. GOVERNMENT CRAWLING AND SCRAPING

Government agencies, from the federal level to local police departments, are already putting information they find online to use. Law enforcement uses social media to anticipate crime,[26] but nearly half of law enforcement agencies have no formal process governing the use of social media for their investigations.[27] This leaves open the possibility of abuse and allows law enforcement professionals to ignore privacy expectations of internet users. The federal government uses data mining to find terrorists by looking for relationships between people and connections between behaviors, and has programs aimed at analyzing "massive" data sets.[28]

Government searches are governed by the Fourth Amendment.[29] Yet whether web crawlers constitute a search under the Amendment is unsettled. There are generally two possible interpretations of the Fourth Amendment's privacy protections: The Third-Party

---

[26] *See* LEXISNEXIS, *supra* note 19, at 3.

[27] *See* LEXISNEXIS, *supra* note 19, at 2.

[28] *See* Jeffrey W. Seifert, *Data Mining and Homeland Security: An Overview*, CONG. RES. SERV. REP. FOR CONG., RL31798, at 26 (April 3, 2008), http://www.fas.org/sgp/crs/homesec/RL31798.pdf.

[29] U.S. CONST. amend. IV. Persistent surveillance online also could have a significant chilling effect on speech. For its First Amendment implications, *see* Karen Gullo, *Surveillance Chills Speech—As New Studies Show—And Free Association Suffers*, ELECTRONIC FRONTIER FOUNDATION (May 19, 2016), https://www.eff.org/deeplinks/2016/05/when-surveillance-chills-speech-new-studies-show-our-rights-free-association.

Doctrine, and a more contextual view of privacy focusing on the amount the surveillance uncovers about a person's life.

### *A. The Fourth Amendment*

The Fourth Amendment's limitation on unreasonable searches applies only to public actors, but it carries great weight in the discussion of online privacy concerns, as the government exercises vast power online to monitor user activity.[30] The Fourth Amendment goes a long distance in shaping the public's perception of their rights in relation to private actors as well, while they are not actually bound by those same constitutional guarantees.

For many years after its conception, courts understood the Fourth Amendment as protecting against a physical invasion of privacy, including a government agent's trespass onto land, or the physical taking of a private citizen's possession.[31] More ephemeral information—like conversations overheard from a location a government agent had a right to be—were granted no protection.[32] It is unclear to what degree trespass may apply to online actions, making it uncertain whether the Fourth Amendment binds government searches online based on a theory of trespass.

Some courts hold that a claim for civil trespass can be sustained based on the use of server resources by a web crawler.[33] In cases where web crawlers used rather small amounts of server resources to search and scrape data from websites, claims against the operators of those web crawlers for trespass have stood.[34] This theory of

---

[30] *See* Glenn Greenwald, *XKeyscore: NSA Tool Collects 'Nearly Everything a User Does on the Internet,'* THE GUARDIAN (July 31, 2013), http://www.theguardian.com/world/2013/jul/31/nsa-top-secret-program-online-data.

[31] *See* United States v. Jones, 565 U.S. 400, 405 (2012) (discussing the history of Fourth Amendment jurisprudence).

[32] *See id.*

[33] *See e.g.*, eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1070 (N.D. Cal. 2000).

[34] *See id.*; *but see* Ticketmaster Corp. v. Tickets.com, Inc., No. CV997654HLHVBKX, 2003 WL 21406289, at *3 (C.D. Cal. Mar. 7, 2003) ("This court respectfully disagrees with other district courts' finding that mere use of a spider to enter a publically available web site to gather information, without more, is sufficient to fulfill the harm requirement for trespass to chattels.").

online trespass is not widely accepted,[35] but it could subject government web crawls to the Fourth Amendment. The architecture of the modern web, which puts nearly everyone's data on someone else's computer through the use of cloud computing, may hamper the use of this doctrine online. The government would not be trespassing on the end user's computer, but onto some company's. In such a case, the end user may never find out, forcing them to rely on others to notify them or to enforce their rights.

In 1967, the U.S. Supreme Court decided *United States v. Katz*,[36] explaining that the "Fourth Amendment protects people, not places."[37] The Court held that a person making a phone call in a phone booth had a reasonable expectation of privacy in his conversation, thus preventing government eavesdropping without a warrant.[38] In later cases, the Court elaborated that a search is unreasonable and violates the Fourth Amendment when the target of the search has manifested an expectation of privacy that society considers reasonable.[39]

### *B. The Third-Party Doctrine*

The Third-Party Doctrine states that there is "no legitimate expectation of privacy in information [one] voluntarily turns over to third parties."[40] A number of cases decided before the creation of the Internet provide for significant government access to records and other information. Applying this doctrine, the courts determined that a number of records held by institutions for or about individuals are unprotected regardless of the use for which they are shared.[41] Courts

---

[35] *See Ticketmaster*, 2003 WL 21406289, at *3 ("[S]cholars and practitioners alike have criticized the extension of the trespass to chattels doctrine to the internet context, noting that this doctrinal expansion threatens basic internet functions (*i.e.*, search engines) and exposes the flaws inherent in applying doctrines based in real and tangible property to cyberspace.").

[36] Katz v. United States, 389 U.S. 347 (1967).

[37] *Id.* at 351.

[38] *Id.* at 353.

[39] *See* United States v. Jacobsen, 466 U.S. 109, 114 (1984).

[40] Smith v. Maryland, 442 U.S. 735, 743–44 (1979).

[41] *See, e.g.*, *Smith*, 442 U.S. at 744; United States v. Miller, 425 U.S. 435, 443 (1976).

held that the Fourth Amendment did not prohibit the government from obtaining information revealed to a third party, even if the information was revealed on the assumption that it will be used only for a limited purpose and the confidence placed in the third party will not be betrayed.[42] This doctrine neatly fits into the *Katz* test, which protects people when they take action to keep their information private. The Third-Party Doctrine adds the presumption that a person can have no legitimate expectation of privacy in shared information.

The impact of the Third-Party doctrine may have been reasonable when it was adopted, but its impact on privacy online is plain and oversized. Online, all of one's activities are shared with a third party. Emails are shared with an email client. The websites one visits are shared with an ISP, and any number of entities that have attached cookies to the browser being used. Everything one does online is shared by the very nature of the Internet; even while browsing alone, some intermediary between one's PC and the server contacted is recording an exchange of packets. As a result, privacy rights are significantly curtailed online. For example, the Electronic Communications Privacy Act, passed in 1986,[43] provides protection against the search and seizure of emails in transit, in storage on a home computer, or stored on what would now be called the "cloud" for 180 days or less. The government must obtain a warrant for such data.[44] For email stored in the cloud for more than 180 days, or opened and stored in the cloud, the government can compel disclosure with only a subpoena.[45] This constitutes less protection than email stored locally, on one's computer (or on paper, in a file cabinet) would get.

### *C. Contextual Privacy*

The views of the Fourth Amendment described above, and the Third-Party Doctrine, assume a black and white view of privacy where any sharing of information, regardless of the purpose,

---

[42] *See Miller*, 425 U.S. at 443.

[43] Electronic Communications Privacy Act of 1986, Pub. L. No. 99-508, 100 Stat. 1848 (codified as amended in scattered sections of 18 U.S.C.).

[44] 18 U.S.C. § 2703(a) (2018).

[45] 18 U.S.C. § 2703(b)(1)(B)(i).

removes any privacy the user could have expected to have in that information. A more nuanced view of privacy is possible, through which internet users would not be denied their privacy based on technical necessities, nor their activities treated as an entirely new realm deserving of a new view of privacy. Instead, the context of the use should be determinative.[46] Just as a patient would be shocked if a doctor shared his information with marketers, but would likely have little issue with that same information being shared with an insurance company or pharmacist,[47] privacy expectations online are contextual.[48] Users share their emails with Google and may expect ads to be shown to them based on the content of those emails, but may not expect those emails to be shared with the government.[49] Under a contextual view, a person's privacy level would depend on the use of the technology.

Of course, applying offline rules to online activities could mean applying the Third-Party Doctrine. But some recent Supreme Court cases might point to a changing view on that issue. In *United States v. Jones*, the Supreme Court unanimously agreed that long-term tracking of a suspect using a GPS device placed on the suspect's car required a warrant.[50] This ruling has significant implications for web crawling. Addressing long-term tracking first, the Court held that it was not reasonable to expect that a government agent would follow someone for a long period of time. Online surveillance and web crawling allow the government to do just that, searching the web and scraping websites for every trace a given user leaves, going back in time as far as any website maintains its data.

---

[46] *See* Helen Nissenbaum, *A Contextual Approach to Privacy Online*, 140 (4) DÆDALUS, J. AM. ACAD. ARTS & SCI. 32, 38 (2011) http://www.amacad.org/publications/daedalus/11_fall_nissenbaum.pdf.

[47] This hypothetical ignores, for the sake of argument, the significant laws that govern the handling of medical information and focuses merely on consumer, or patient, expectations.

[48] Nissenbaum, *supra* note 46, at 38.

[49] Additionally, in the particular case of email, the change in how email is used since the passage of ECPA and the routine storage of large numbers of emails and other documents in the cloud, rather than on home computers, bolsters the argument that users do expect a different amount of privacy than ECPA provides, at the very least.

[50] United States v. Jones, 565 U.S. 400, 404 (2012).

A concurrence by four justices rejected the majority's trespass-based approach and determined that a reasonable person did not expect to be tracked with a GPS unit over a long period of time, which in this case, was about a month.[51] The reasoning of the four concurring justices, adopting a new approach to apply to persistent, long-term tracking which was either impossible or prohibitively expensive in the past, may signal a coming change in how such cases are decided. Such a view may even lead to a significant curtailing, if not the end, of the Third-Party Doctrine.

Justice Sotomayor, in her own concurrence, expressed concern over the application of trespass in the electronic age given that many forms of surveillance require no trespass. For instance, tracking the GPS chip in a suspect's phone, rather than placing one somewhere on his person or possessions.[52] Justice Sotomayor was explicitly worried about electronic surveillance and went as far as suggesting that the Third-Party Doctrine be reconsidered. She said the approach was "ill suited to the digital age, in which people reveal a great deal of information about themselves to third parties in the course of carrying out mundane tasks."[53] As one scholar put it, "all communications over the Internet . . . are stored for various lengths of time on third party servers or Internet service providers."[54] Justice Sotomayor cited *Katz* for the proposition that "what [a person] seeks to preserve as private, even in an area accessible to the public, may be constitutionally protected."[55] Further, computers, including those online or in the cloud, are routinely used to hold the sorts of documents, photographs, and other private matters that were previously kept in the home.[56] Without changes to the Third-Party Doctrine, these documents would lose protection merely because of where they are stored.

---

[51] *Id.* at 418 (Alito, J., concurring).

[52] *Id.* at 413 (Sotomayor, J., concurring).

[53] *Id.* at 417.

[54] Monu Bedi, *Facebook and Interpersonal Privacy: Why the Third Party Doctrine Should Not Apply*, 54 B.C. L. REV. 1, 2 (2013).

[55] *Jones*, 565 U.S. at 418 (citing Katz v. United States, 389 U.S. 347, 351–52 (1967)).

[56] *See* Katherine J. Strandburg, *Home, Home on the Web and Other Fourth Amendment Implications of Technosocial Change*, 70 MD. L. REV. 614, 654–55 (2011).

In *Riley v. California*, the Supreme Court discussed how searches of cell phones can reveal far more than just one sort of information contained in them would otherwise reveal.[57] In this case, the government searched a cell phone incident to arrest.[58] The Court reaffirmed that searches of cell phones under this authority must occur to protect officer safety or to preserve evidence, and otherwise require a warrant or exigent circumstances.[59] However, recognizing the difference between collecting large and small amounts of information has clear implications for government use of web crawlers.

Although neither *Jones* nor *Riley* addressed online surveillance specifically, it seems clear that long-term surveillance, or surveillance that covers a wide variety of information (and perhaps even information shared online in at least some contexts) may not be completely unprotected under the Fourth Amendment. These cases drew a line based on the amount of data collected; they alleged that when the government collects enough data, even if it is public, the nature of the collection can change and violate a persons' privacy.

Scholars have suggested new ways to apply the Fourth Amendment online in a way that would protect the privacy of those who share information online. One way is to protect content, while allowing the government to collect non-content information.[60] This was proposed as being similar to the inside/outside distinction applied in physical space, in which people have a greater degree of protection under the Fourth Amendment inside, in private spaces, than they do outside, in public.[61] This is also similar to the

---

[57] *See* Riley v. California, 134 S. Ct. 2473, 2489 (2014).

[58] *Id.* at 2482.

[59] *Id.* at 2483.

[60] *See* Orin S. Kerr, *Applying the Fourth Amendment to the Internet: A General Approach*, 62 STAN. L. REV. 1005, 1029 (2010). Non-content information, or metadata, is information "related to identity, location, and time." *Id.* at 1018. Metadata could feasibly include email addresses, account names, IP addresses, or other similar information. *See also* Chris Conley, *Metadata: Piecing Together a Privacy Solution*, ACLU OF CALIFORNIA (Feb. 2014), https://www.aclunc.org/sites/default/files/Metadata%20report%20FINAL%202%2021%2014%20cover%20%2B%20inside%20for%20web%20%283%29.pdf.

[61] Kerr, *supra* note 60, at 1009.

protections currently applied to post mail and telephone calls,[62] but may draw critics based on the revealing nature of metadata.[63]

Alternatively, one could apply Fourth Amendment protections online based on the "structure of the particular technology" and "the particular uses to which an individual puts the technology."[64] Under such an approach, password protected information stored in the cloud would be protected, even if it were non-content information, just as if it were held in a filing cabinet in one's home.[65] Determining how to deal with social media is difficult under this approach, but could be determined based on the amount of control the user maintains over access to the information, even if the owner of the platform has access for certain purposes.[66] The court could ask if "assuming privacy settings are optional, [the 'resident'] chose privacy settings that would support a finding that his [social media sites are] sufficiently restricted that they are not readily available to the general public."[67] Just as in determining whether to treat a physical space as a residence, courts should not inquire too closely into the specific uses an individual chooses to make of an online social space; an individual does not have a lesser basic expectation of privacy against the government in their home simply because they have frequent parties or have a large number of guests.[68]

Finally, the Fourth Amendment could be read to protect certain "structural privacy rights."[69] Acknowledging that prior to certain technological advancements, some forms of surveillance were too expensive to employ, the courts should strive to maintain protections at that level. For example, while following a given person was once prohibitively expensive, one can now be followed electronically with the use of the GPS chip in one's phone. A rule designed to

---

[62] *Id.* at 1019.

[63] *Id.* at 1032.

[64] Strandburg, *supra* note 56, at 659–60.

[65] *Id.*

[66] *Id.* at 661–62.

[67] *Id.* at 663 (citing Crispin v. Christian Audigier, Inc., 717 F. Supp. 2d 965, 991 (C.D. Cal. 2010)).

[68] *Id.*

[69] *See* Kevin S. Bankstson & Askan Soltani, *Tiny Constables and the Cost of Surveillance: Making Cents Out of* United States v. Jones, 123 YALE L.J. 335, 339 (2014).

protect a structural privacy right would use the Fourth Amendment to impose legal costs where there were once economic costs.[70]

## IV. PRIVATE SECTOR CRAWLING

The private sector may have many uses for crawling and for scraped data beyond those discussed above. Companies can use them to gather information on their customers' views on certain products they've purchased. They can gather information about pricing on their competitors' websites. They could also be used to gather significant amounts of information on their customers from personal blogs, social media sites, forums, and other websites where users may talk about or otherwise make their identity or their preferences known. This could allow companies to gather large dossiers of sensitive information with few, if any, rules about what can be gathered, when and where it can be gathered from, along with generally weak rules about the storage of information. This section will discuss the case law applicable to corporate use of web crawlers and the policy implications of corporate use. Some sectors of the U.S. economy are governed by industry-specific privacy regulations.[71]

### *A. Trespass*

In *eBay v. Bidder's Edge*, a California district court was faced with determining whether Bidder's Edge, an auction aggregation site, could crawl eBay's website, scrape information on bids, and provide search results to its own users.[72] The court held that such unpermitted crawling amounted to trespass, and ordered an injunction to stop Bidder's Edge from continuing its crawling and scraping of eBay.[73] The court came to this decision even though Bidder's Edge used very little of eBay's server resources (a couple of percent, at most), and did not damage the property, though it did

---

[70] *Id.*

[71] These privacy regulations will be discussed where applicable, but they are relatively narrow in scope and are largely outside the scope of this paper.

[72] eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1067 (N.D. Cal. 2000).

[73] *Id.* at 1069–70.

prevent eBay from using a small percent of server resources for other uses.

Another California court attempted to apply this "ancient common law action to the modern age."[74] Prior courts held that "mere invasion or use of a portion of the web site by a spider is a trespass (leading at least to nominal damages), and that there need not be an independent showing of direct harm either to the chattel (unlikely in the case of a spider) or tangible interference with the use of the computer being invaded."[75] The *Ticketmaster* court, however, required a showing that the computer being crawled be adversely affected by the use of the spider, rejecting that "mere use of a spider to enter a publicly available web site to gather information, without more, is sufficient to fulfill the harm requirement for trespass to chattels."[76]

The California Supreme Court dealt with a similar issue where a former company employee sent a number of emails to his former coworkers' corporate email accounts.[77] Here, a number of emails were sent to employees, who were given the choice to opt out of receiving the emails.[78] Intel argued that it deserved an injunction against the sending of those emails, as the emails were a trespass on its server that ate up server and human resources (time spent replying, setting up filters, etc.).[79] However, the court declined to find a trespass, as California law required some damage to the property. Here, there was no allegation that the emails impaired the functioning of Intel's computers, and the emails were allowed to be sent.[80]

Courts have come to vastly different conclusions about whether trespass applies online, and have made some important points in doing so. First, it is important to note that *Intel v. Hamidi* depended on the definition of trespass, a common law concept that can differ

---

[74] Ticketmaster Corp. v. Tickets.com, Inc., No. CV997654HLHVBKX, 2003 WL 21406289, at *3 (C.D. Cal. Mar. 7, 2003).

[75] *Id.* (noting the criticism of extending trespass to the internet).

[76] *Id.* (disregarding the work load performed by Ticketmaster's servers to accommodate Tickets.com's crawlers).

[77] Intel Corp. v. Hamidi, 71 P.3d 296, 299 (Cal. 2003).

[78] *Id.*

[79] *Id.* at 300.

[80] *Id.* at 311.

from one jurisdiction to another. Second, it is unclear what the definition of damage is when applied to the use of a server. One court found that merely using server resources was enough to find damage, while another found that a minimal use that did not affect the operation of the computer at issue was not enough for a court to find damage.[81] It is unclear, based on these opinions, whether merely using a computer's resources constitutes damage, and if not, how much of a computer's resources must be used for a court to find it was damaged. It is also unclear what sort of warnings are required to make it known that a crawler is unwelcome. *Intel v. Hamidi* did not address the issue in-depth,[82] and eBay notified Bidder's Edge in multiple ways that their crawlers were unwelcome.[83] Would merely having a robots.txt header forbidding crawling or posting it in a website's ToS be enough? If *any* use of server resources without permission is a trespass, then how can the operator of a crawler find out what is in a target website's robots.txt header or ToS without crawling? The common law cause of action of trespass does not provide a rule clear enough for the operators of web crawlers to follow, and leaves enforcement largely up to websites, not end users whose data is actually at issue. It is not enough to ensure user privacy from web crawlers only when it is desired.

### *B. The Computer Fraud and Abuse Act*

The Computer Fraud and Abuse Act of 1986 (CFAA)[84] protects computers from unauthorized access and from access that exceeds authorization.[85] The law provides for both criminal and civil penalties.[86] At times, courts have addressed whether unauthorized crawling and scraping can violate the CFAA. Because the CFAA

---

[81] *Compare eBay v. Bidder's Edge, Inc.*, 100 F. Supp. 2d 1058 *with* Ticketmaster Corp. v. Tickets.com, Inc., No. CV997654HLHVBKX, 2003 WL 21406289, at *3 (C.D. Cal. Mar. 7, 2003).

[82] *Id.* at 300. In fact, Intel did not appeal to Hamidi to stop sending the messages, but merely attempted to block the receipt of them by Intel employees.

[83] eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058, 1062 (N.D. Cal. 2000).

[84] Computer Fraud and Abuse Act, 18 U.S.C. § 1030 (2018).

[85] *Id.*

[86] *Id.* § 1030(c).

was passed in 1986, it does not incorporate web crawlers into its provisions. It is not clear how the law would apply to such software, as the following cases illustrate. Further, courts have been divided over how the CFAA should be applied outside of the limited case of web crawlers.[87]

In *EF Cultural Travel BV v. Explorica*, the First Circuit was tasked with determining whether scraping a website violated the CFAA. Its determination of whether access was unauthorized in this particular case is outside the scope of this paper, as it hinged on a confidentiality agreement signed by a former employee of the company whose website was scraped, and not on an html header, ToS, or other commonly used means of signaling a desire not to be crawled or scraped.[88] However, the court also looked at whether the scraping met the damage or loss requirements of the CFAA. The court found that EF Cultural Travel had suffered a loss due to Explorica's scraping, under a theory that Congress had intended loss "to target remedial expenses borne by victims that could not properly be considered direct damage caused by a computer hacker."[89] Because EF Cultural Travel had been forced to take "diagnostic measures" to "assess whether their website had been compromised,"[90] they had suffered a loss. Though EF Cultural Travel suffered no physical damage, the court determined that Congress, by specifying that either damage or loss would enable recovery under the CFAA, had intended that no physical damage was necessary.[91] However, nine years later, the District Court of Maryland held that for lost revenue to qualify as a "loss" under the CFAA, the unauthorized access in question must have caused an interruption of service.[92] Other courts have declined to follow that definition.[93]

---

[87] *See* Orin S. Kerr, *Norms of Computer Trespass*, 116 COLUM. L. REV. 1143, 1143–44 (2016).

[88] EF Cultural Travel BV v. Explorica, Inc., 274 F.3d 577, 582 (1st Cir. 2001).

[89] *Id.* at 585 (citing *In re* DoubleClick Inc. Privacy Litig., 154 F. Supp. 2d 497, 521 (S.D.N.Y. 2001)).

[90] *Id.* at 584 & n.17.

[91] *Id.* at 585.

[92] CoStar Realty Info., Inc. v. Field, 737 F. Supp. 2d 496, 513 (D. Md. 2010).

[93] *See, e.g.*, Am. Family Mut. Ins. Co. v. Gustafson, No. 08–cv–02772–MSK,

In another case, *AOL v. LCGM*, the court held that LCGM violated the CFAA by sending bulk email to AOL subscribers in violation of AOL's ToS[94] and by collecting those email addresses in violation of the same ToS.[95] Again, LCGM caused AOL to incur technical costs as a result of their actions, impaired the functioning of AOL's network, and damaged AOL's goodwill.[96]

Over the years, courts have operated under a number of different rules regarding when the CFAA applies. However, it seems clear that a web crawler visiting a target website, using its resources, and scraping it for data, could violate the CFAA. Web crawlers can certainly operate in violation of an html header or of a ToS,[97] and they also use resources of the servers they contact, which could cause a service disruption. Consequently, website operators wishing to keep crawlers away from their site must expend money and resources responding to such visits.

Nevertheless, in a recent case the Northern District of California found there was likely no violation of the CFAA in a suit brought by LinkedIn against hiQ, which scraped LinkedIn for publicly accessible data in violation of LinkedIn's ToS.[98] The court distinguished previous cases,[99] finding a CFAA violation in similar circumstances, while noting that unlike previous cases, hiQ was scraping public data rather than password protected parts of

---

2011 WL 782574, at *4 (D. Co. Feb. 25, 2011) (finding that "loss" is limited to "cost[s]" and to "any revenue lost, cost incurred, or other consequential damages . . . incurred because of interruption of service," and holding that lost revenue was not a "loss"); First Fin. Bank, N.A. v. Bauknecht, 71 F. Supp. 3d 819, 851 (C.D. Ill. 2014) ("[T]here are two categories of statutory loss: expenses incurred while responding to or investigating a violation, and costs incurred, or revenue lost, because of a service disruption.").

[94] Am. Online, Inc. v. LCGM, Inc., 46 F. Supp. 2d 444, 450 (E.D. Va. 1998).

[95] *Id.* at 450–51.

[96] *Id.* at 451.

[97] *See, e.g.*, Kerr, *supra* note 87, at 1165–67 (noting that some scholars do not think that ToS should be binding on web users, as they are rarely read, hard to understand, and better understood as limits on liability than as limits on who can use the website).

[98] hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099, 1108 (N.D. Cal. 2017).

[99] *Id.* (citing United States v. Nosal, 844 F.3d 1024, 1038 (9th Cir. 2016) & Facebook, Inc. v. Power Ventures, Inc., 844 F.3d 1058, 1067 (9th Cir. 2016)).

websites.[100] The court explained that, unlike in *United States v. Nosal* or *Facebook v. Power Ventures*, where "unauthorized intruders reached into what would fairly be characterized as the private interior of a computer system not visible to the public,"[101] the scraping at issue here was publicly available, without a password, and this put it outside Congress' intent in passing the CFAA to prevent hacking.[102] Further, the court reasoned that applying the CFAA in the way LinkedIn suggested "would have sweeping consequences well beyond anything Congress could have contemplated," potentially creating criminal liability for "merely *viewing* a website in contravention of a unilateral directive from a private entity . . . effectuating the digital equivalence of Medusa."[103] The court also discussed how to apply the concept of trespass to online domains, determining that social norms tell us the Web is "inherently open," and that the CFAA's bar on "access without authorization" probably does not apply to publicly available portions of a website.[104] The court awarded hiQ a preliminary injunction barring LinkedIn from preventing hiQ's scraping activity on their website.[105]

### *C. Overview of Private Sector Use*

Private sector corporations are subject to significant restrictions on what and when they can crawl. Unlike the restrictions on the government, these restrictions are not theoretical, though they are hardly clear-cut. It seems that corporate operators of web crawlers may need to abide by the desires of websites to not be crawled, whether that preference is made known in a robots.txt header, a ToS, or otherwise. However, this is dependent on the ability and willingness of websites to use litigation to stop crawlers from operating on and scraping their website, leaving smaller websites and users in a jam.

---

[100] *Id.* at 1109.
[101] *Id.*
[102] *Id.*
[103] *Id.* at 1110.
[104] *Id.* at 1111 (citing Kerr, *supra* note 87, at 1162).
[105] hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099, 1120 (N.D. Cal. 2017).

These controls on private operators of web crawlers are available only to the operators of a website. Individual users cannot ensure their data is not crawled or scraped, and must rely on the operators of the websites they use to maintain their privacy against crawlers. Given how vague case law is on the subject, it is unclear whether users or websites can rely on these protections to keep their data private and out of corporate databases. Many websites and users may be unable to protect themselves, and some websites may find it is in their interest to allow crawlers to scrape their data, regardless of some of their users' wishes.

For example, web forums may lack the resources and money to defend their users' information from those who wish to scrape it. While some forums are quite large, most are small and likely lack the technical, monetary, and legal resources to stop an organization that insists on ignoring their calls to refrain from crawling and scraping. These forums may be quite interested in protecting their data; forums often host discussions on personal issues, including those of sex, medical conditions, and others, and have a reputation that they wish to maintain among their users. However, they often do not monetize this data beyond serving ads to those who read or post. This limits their resources and how valuable that data is to the forum; they lose no value if another group holds the same data. These sorts of forums may not be willing or able to protect their users' privacy and users have no way of signaling their desire not to have their posts crawled, and suffer even more from a lack of resources. Other websites, like Twitter, do monetize the data they collect by limiting the ways that data can be culled from their service and charging users to access the full archive of tweets.[106]

Social networks collect even more data than forums, and this data is perhaps more sensitive and specific than that people post on

---

[106] *See* Juliette Garside, *Twitter Puts Trillions of Tweets up for Sale to Data Miners*, THE GUARDIAN (Mar. 18, 2015), http://www.theguardian.com/technology/2015/mar/18/twitter-puts-trillions-tweets-for-sale-data-miners; *Twitter firehouse vs. Twitter API: What's the Difference and Why Should You Care?*, BRIGHTPLANET (June 25, 2013), http://www.brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care/. *See also* @raffi, *Twitter #DataGrants Selections*, TWITTER (Apr. 17, 2014), https://blog.twitter.com/2014/twitter-datagrants-selections (explaining that Twitter does supply free access to its complete archive of tweets to select universities through its #DataGrants program).

forums. But like forums, social networks have a reputation to protect, and the larger ones may have significant resources and a desire to keep whatever information they have to themselves, and monetize it as they see fit. For example, Facebook collects, and reveals, large amounts of data about its users. It uses the data to make recommendations, displays news stories of potential interest, and shows advertisements based on the information scraped. In 2018, amid a media firestorm, Facebook's CTO confirmed that a private company, Cambridge Analytica, surreptitiously scraped data from 87 million users.[107] The firm reportedly collected the Facebook profiles in order to target voters during the 2016 U.S. Presidential election.[108] This incident focused international attention on the risk of crawlers deployed by third parties harvesting detailed personal data found on proprietary social networks.

## V. ACADEMIC USE

Crawlers also have potential for academic researchers in social science, computer science, and other fields. Internet research has greatly expanded the methods for social analysis used by researchers. Now, in addition to traditional surveys, researchers can collect vast amounts of data from online communities, social media, and various websites to answer questions on topics such as youth attitudes, demographic change, or political beliefs.

In the same way that the government or corporations may use web crawlers to collect sensitive data that users meant to keep private, researchers may collect significant data on a much wider array of issues of noncommercial general inquiry. While searching for private, closely held beliefs and ideas can lead to valid findings, researchers in academic institutions are bound by the same laws that govern the private sector and have additional institutional controls over their research.

---

[107] Anne L. Washington, *Facebook math: How 270,000 became 87 million*, DATA & SOCIETY: POINTS (April 11, 2018), https://points.datasociety.net/facebook-math-how-270-000-became-87-million-bd8cf1009b32.

[108] Kevin Granville, *Facebook and Cambridge Analytica: What You Need to Know as Fallout Widens*, THE NEW YORK TIMES (Mar. 19, 2018), https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

The CFAA arguably stands in the way of academics who want to use crawlers. Researchers may seek to deploy web crawlers and other bots to gather and analyze data for basic and applied research publications adding to literature of their disciplines. The tension surrounding this use is not theoretical. In 2017, University of Michigan Professor Christian Sandvig, his academic colleagues, and the news organization First Look Media Works, intended to conduct research on online discrimination using methods including web crawlers where such conduct is prohibited by the CFAA. The ACLU filed a lawsuit on their behalf against the U.S. Attorney General over the CFAA's criminalization of such research activities according to the website's ToS.[109] The plaintiffs are concerned that conducting their research with crawlers, which they allege will cause no harm to the websites they study, will expose them to significant criminal liability. The case has not yet been decided on the merits, but they have been allowed to move forward with an as-applied challenge to the CFAA on the Free Speech and Free Press Clauses of the First Amendment. Even if their case is successful, the website ToS will remain in force and they may be prohibited from accessing the websites themselves or be subject to civil actions.

Academics performing studies have more oversight on their research than some other actors. Institutional Review Boards (IRBs) are tasked with reviewing and approving proposed human research by academics. IRBs are supposed to ensure that researchers obtain informed consent from their subjects and do not expose them to undue risk of harm.

However, there are number of problems with the IRB process. First, they often take a long time to complete their reviews (often months), keeping them slightly behind the newest technology. They also may not necessarily understand the problems associated with collecting data online; while using publicly available data posted on the web may not appear to be human subjects research, such data use clearly can have significant impact on the lives of those who posted it. Finally, many researchers use "found" data, or data that has been collected by another entity, which is either publicly

---

[109] *See* Sandvig v. Sessions, No. 16–1368 (JDB), 2018 WL 1568881, at *4-5, (D.D.C. Mar. 30, 2018). The CFAA, for example, also acts upon academic users of web crawlers.

available online or given to them by a private company, without further review.[110] This allows researchers to avoid institutional review even when they are subjecting the data to new analysis and may uncover novel findings about those who posted the data online. Such use creates another point of failure where personally identifiable information can be revealed or data can be leaked. Considering the possible problems with avoiding review in this way is made more important in light of recent calls for researchers to open up the data they use in their research and to share it with others in their field.[111]

Academic researchers need clearer rules about mandatory review of the analyses they wish to perform on this sort of data, even when it is collected by another entity. Academic actors collect information and perform studies on topics that are just as sensitive as the projects carried out by the government. They study religion, sex, gender, and a host of other topics, many times focusing on vulnerable or disenfranchised populations. Institutions reviewing this sort of research need to ensure that the studies they produce are conducted with respect for the privacy of those using the internet and that the data collected is handled and saved responsibly.

## VI. APPLICATION

Given this state of affairs, users may enjoy some degree of privacy online, even in the information that they post publicly. However, the existing laws and guidelines governing the use of web crawlers to gather information on the web are inadequate to the task

---

[110] *See* 45 C.F.R. § 46.101(b)(4) (exempting from the human research subjects policy "Research, involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.").

[111] *See, e.g.*, Paige Shaklee, *New Data Journal Lets Researchers Share Their Data Open Access*, ELSEVIER CONNECT (Sep. 9, 2014), https://www.elsevier.com/connect/new-data-journal-lets-researchers-share-their-data-open-access ("[E]ach piece of data that has been carefully and thoughtfully gathered has value. Often, you don't know what future value that data will have until you've shared it with colleagues in the scientific community.").

of protecting privacy interests. While the courts have not dealt with government surveillance using web crawlers, a wide search could turn up enough information, in aggregate, to create a search subject to the Fourth Amendment. Just as tracking a person with a GPS unit for 30 days reveals much about that person's life, so could crawling and scraping enough data about a particular person. Such searches threaten to reveal nearly everything about a person's life without the knowledge of those being searched.[112] Law enforcement also recognizes that using online material for policing purposes requires walking a fine line. The Bureau of Justice Assistance produced a report recommending that police departments institute policies governing when such tools can be used, what authorization is needed, and how collected data should be stored.[113]

A similar expectation of privacy exists against privately operated web crawlers, though this expectation is largely enforceable only by the website hosting the information, not the end user. While online trespass is not widely accepted as a good idea among the legal community, and the CFAA was not aimed specifically at protecting from this kind of harm, these bodies of law do provide some protection against robot searches. Such crawls, if unwanted, could create a private cause of action against those operating the web crawlers, though there are practical concerns to enforcing such a prohibition on crawling.

Beyond the legal norms discouraging unwanted crawling and scraping of data from websites, ethical and social norms are in place. Facebook, whose founder once said that privacy was no longer a social norm, has changed its sharing default from "public" to "friends."[114] Eighty-six percent of internet users have taken some

---

[112] These could reveal locations from check-ins and photos on social networks, opinions about politics, social movements, and literature, names of friends and acquaintances, product reviews on online marketplaces, and more.

[113] *Developing a Policy on the Use of Social Media in Intelligence and Investigative Activities: Guidance and Recommendations*, GLOBAL JUSTICE INFO. SHARING INITIATIVE ADVISORY COMM., at 9 (Feb. 2013), https://it.ojp.gov/documents/d/Developing%20a%20Policy%20on%20the%20Use%20of%20Social%20Media%20in%20Intelligence%20and%20Inves....pdf.

[114] *See* Molly Wood, *Facebook Generation Rekindles Expectation of Privacy Online*, N.Y. TIMES: BITS (Sept. 7, 2014), http://bits.blogs.nytimes.com/2014/09/07/rethinking-privacy-on-the-

step to remain private online, and sixty-eight percent say that stronger laws are needed to protect people's online privacy.[115] People attempt to guard their identity, keep information from specific people or organizations, and care quite strongly that they control who has access to much of their information.[116]

To ensure that internet users' privacy is maintained, more work is needed to put in place strong administrative and legal protections. At the moment, it is unclear how the law applies to web crawlers in all jurisdictions. Private sector actors, including academic institutions, have weak controls on their use of these tools. More accountability is needed, and clearer rules need to be put in place to ensure that web crawlers are not abused and internet users do not suffer undue harm. The remainder of this paper will discuss some of the policy questions that need to be considered while crafting these rules.

## VII. POLICY DILEMMAS

Internet users have certain expectations about their use that web crawlers may confound. Certain social norms exist surrounding use of the Internet and particular websites on it. For example, when users post an update on Facebook, they expect that post is for the use and enjoyment of their friends. Though it may be available to the public, most people are unlikely to think that their posts will be scrutinized and used to profile them.[117] Further, many websites have rules prohibiting web crawling, contributing to the belief that people's data will not be scooped up by a bot sent on a mission to find any data that it can. Government, corporate, and university web crawling

---

internet/?_r=0.

[115] *See* Lee Rainie et al., *Anonymity, Privacy, and Security Online*, PEW RES. CTR. (Sept. 5, 2013), http://www.pewinternet.org/2013/09/05/anonymity-privacy-and-security-online/.

[116] *Id.*

[117] *See* Motahhare Eslami et al., *"I Always Assumed that I Wasn't Really That Close to [Her]": Reasoning About Invisible Algorithms in News Feed*, 33 PROC. ANN. ASS'N FOR COMPUTING MACHINERY (ACM) CONF. ON HUMAN FACTORS COMPUTING SYSS. (CHI 2015) 153 (2015), http://www-personal.umich.edu/~csandvig/research/Eslami_Algorithms_CHI15.pdf.

shatters that expectation.[118] It allows large organizations to build a comprehensive profile on any person or organization it would like to, at very low cost to those operating web crawlers.

### *A. Metadata*

Crawlers allow for the compilation of a significant amount of metadata about users. This metadata can be extremely revealing, is often unprotected, and may not be protected from government search under the Fourth Amendment. With some effort, metadata from anonymous accounts could be linked to a real identity, meaning that users could not escape being tracked by using an alias or username not plainly associated with them. A person's religious views, medical status, or other personal information could be determined just from viewing metadata.

This information could be embarrassing, used against people in courts or among the public, and could be data that a person never wanted linked back to their real identity. Using web crawlers to collect and index this sort of data could thwart all of those expectations.

### *B. Exclusions and Bias*

Crawlers do not, and perhaps cannot, search everything. They will inevitably miss information, fail to search some websites, or mistakenly believe that some information is not relevant to its search and fail to collect it. As with all other methods of data collection, some people and data will be excluded from the searches conducted by crawlers. What this means for those operating web crawlers is not entirely known. In the context of the government, it means that searches for criminals will never be perfect. For corporations or researchers, it means that searches designed to study a given community will miss people, and fail to provide a full picture. This could bias any resulting conclusions drawn from such data, and require that those directing searches consider how inclusive their search will be and ways to correct for such exclusion bias.

Searches conducted with crawlers will suffer from more

---

[118] Though, after the release of the Snowden documents, people may be more aware of the surveillance they are subject to online. *See also* Esposito et al., *supra* note 20.

traditional forms of bias. Just as someone drafting questions for an opinion poll may chose words that push people towards a certain answer, programmers may choose search terms, or construct their algorithms in such a way that their bots are drawn to certain types of data, and hence certain types of answers. This also leaves open the possibility that the searching organization may miss someone, mistakenly associate someone with an act, or may make improper conclusions on which policy will be based.[119]

Not all of these are strictly privacy problems. The fact that someone was not found by a crawler is surely a good thing for their privacy, but may be bad for public policy. At the same time, invading peoples' privacy imperfectly leaves open the possibility that action will be taken against people who, in truth, should be left to lead their lives in peace. Controls need to be put in place based on realistic abilities of web crawlers in finding information to ensure that does not happen.

### *C. Data Security*

Collecting large amounts of data makes one a target for hackers and opens the possibility of data leaks. As discussed above, this data can be sensitive and can paint a detailed picture of a person's life. Government agencies have not yet found practical ways to secure their data, and have publicly failed to do so.[120] Before they embark on additional data collection initiatives, any actor needs to ensure that it can keep the information it does collect safe. This means strong access controls, employing encryption to protect the data, ensuring that employees practice good 'cyber hygiene,' that computers are regularly updated, and that steps are taken against unauthorized outside access.

---

[119] *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1267 (2008).

[120] *See The OPM Data Breach: How the Government Jeopardized Our National Security for More than a Generation*, U.S. HOUSE OF REPRESENTATIVES COMMITTEE ON OVERSIGHT AND GOVERNMENT REFORM MAJORITY STAFF REPORT (Sep. 7, 2016), *available at* https://oversight.house.gov/wp-content/uploads/2016/09/The-OPM-Data-Breach-How-the-Government-Jeopardized-Our-National-Security-for-More-than-a-Generation.pdf.

### *D. Future Uses*

Just as internet users probably do not expect their data to be collected and used for government purposes when they post on Facebook or one of the many forums that exist, they certainly do not expect their data to be used in the future for purposes not yet imagined. Data storage is increasingly inexpensive and allows for the long-term storage, and therefore the long-term use, of collected data.[121] While many of the things that people post online fade in their ability to cause embarrassment or harm with age, many do not, and some may in fact end up more potent in that regard.

If organizations are to collect data with web crawlers, even in a limited scheme, it must consider whether it plans to maximize the amount of data it collects, over-collecting and storing indefinitely, or minimize its data, discarding it as it is used or after a given time period, during which it is put to no use. Data should, in all cases, be minimized to protect the privacy of internet users, who should not have to worry that decades after posting, their youthful indiscretions will haunt them because a government crawler saved a post.

### *E. Unfair or False Light, Undue Harm, and False Positives*

Related to some of the other concerns listed here, data could be used to paint some internet users in an unfair or false light. Failing to fully collect data about people, or using only part of the data collected, could make a person look bad for failure to consider context or the full picture. This sort of risk can be reduced by controlling how data will be used, who has access to it, and how long it is kept. Use of this data could cause severe harm to some internet users, and may point a guilty finger at innocent users. Organizations employing web crawlers to collect data should consider what level of certainty is required before they can employ their data. There should also be procedural hurdles before such data

---

[121] Lucas Mearian, *CW@50: Data Storage Goes from $1M to 2 Cents per Gigabyte (+Video)*, COMPUTERWORLD (Mar. 23, 2017), https://www.computerworld.com/article/3182207/data-storage/cw50-data-storage-goes-from-1m-to-2-cents-per-gigabyte.html (noting that from the year 2000 to 2017, the cost of a gigabyte stored on a disk drive has dropped from $7.70 to $0.02).

can be used; just as the criminal justice system is governed by proof beyond a reasonable doubt, programs using data from web crawlers need similar, if less lofty, standards governing their actions.

### *F. Misuse of Data*

There is also the possibility of deliberate misuse of data. Individual employees may use their resources to further their own ends, or simply for entertainment. Proper access controls and good security can significantly reduce the risk of this and protect internet users swept up by web crawlers from significant embarrassment and possibly serious harm.

### *G. Vulnerable Populations*

Many vulnerable, hidden, or marginalized populations use the online technologies to communicate to find support.[122] Sometimes this is done in the open on Twitter, in forums, or through other clients that keep records of their discussions on the open, searchable web. Government agencies may decide some of these populations need to be watched, either for their own safety or the safety of others. This could do significant damage to such communities, causing them to disband after discovering they are under surveillance, or subjecting them to discrimination because of what is found in discussions they never intended for outsiders.

### *H. Chilling Speech*

Finally, government surveillance can have the effect of chilling speech. Those who know the government is crawling the web to record conversations, metadata, and other information may choose not to have conversations or not to go online in the first place. This has significant social costs, and the government should consider the public, civic, and social goods that the internet fosters before it takes actions that could hinder those acts that make the internet so

---

[122] *See e.g.*, UNHCR, *Connectivity for Refugees*, www.unhcr.org/innovation/connectivity-for-refugees/ (last visited May 8, 2018); *see also* THE ECONOMIST, *Phones are now Indispensable for Refugees*, Feb. 11, 2017.

valuable.

## VIII. HOW TO TREAT ROBOTS ONLINE

The internet is undoubtedly an open place that users should be able to surf free of fear from legal action over trespass from website operators with extreme ToS or other usage controls.[123] However, the widespread use of web crawlers to collect information may confound the expectations of many internet users who do not have full knowledge of how the internet works and what bots are capable of. People may understand that their comments will persist, and may be linked to their identity, but the abilities enabled by bots go beyond the risk that a stray comment or account will be linked to a real identity.

Internet users take part in online communities with expectations as to how those communities operate and how their contributions will be maintained. They largely assume that humans and the service they are using will read their posts and review their activity, not some outside party. Website owners also have expectations that they will be able to monetize the data they collect, and that data will not be taken without compensation.

Web crawlers confound these expectations by giving anyone the ability, with relatively few resources, to collect huge amounts of information posted online. While this may threaten business models, it also threatens the assumption of relative obscurity that many users depend on when they partake in online forums. The scale on which robots, and not humans, can collect information, is the relevant consideration in determining whether websites should be allowed to control access by robots.

Web crawlers may require different handling. Website owners should be able to count on robots.txt to guide robots that access their webpages. This would allow website owners to make it clear which pages robots can access and perhaps, how often, and is a clear line for courts trying to apply trespass or other authorized access laws to the internet.

The analysis is not entirely dissimilar from the analysis applied

---

[123] Kerr, *supra* note 87, at 1162.

by the court in *hiQ v. LinkedIn*.[124] While the court there proposed that the situation is more similar to a shop that has "displayed a sign in its storefront window visible to all on a public street and sidewalk," where "it could not ban an individual from looking at the sign and subject such person to trespass for violating such a ban,"[125] the analogy ignores the fact that online, one cannot look at a shop without entering it. A more apt analogy may be if someone walked into that same shop with a scanner, and saved digital copies of its wares for later reproduction and use. Nevertheless, robots.txt could be seen as analogical to a shop owner restricting the manner and scope of access to a physical store.

Enabling website owners to undertake civil actions for violations of their robots.txt restrictions acts similarly to trespass norms; owners can decide who is allowed on to their property, and for what purposes. This solution is not perfect for a number of reasons. It leaves owners of websites in charge of determining and enforcing the wishes of their users, and leaves some web crawler users who people might want to allow to have their information, such as researchers, without that access. This can occur in cases where website owners are indiscriminate in their rulemakings or limit access by corporate entities that publish databases used by researchers. Limiting the rules specifically to bots also addresses some of the possible negative outcomes of applying the CFAA to scraping that the court noted in *hiQ* —namely, consequences ranging from racial or gender discrimination to illiberal political outcomes.[126]

However, owners of websites are far more likely to be responsive to users' wishes than the more detached third parties operating web crawlers. Additionally, those who want access to the information currently gathered with web crawlers can negotiate for it, something that already happens with many websites like Twitter.[127] This leaves website owners in control of who can gather

---

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

[125] *Id.* at 1112–13.

[126] *Id.* at 1110.

[127] *See, e.g.*, Barry Schwartz, *Google Confirms New Experiment with Twitter in Search Results*, SEARCH ENGINE LAND (May 4, 2015),

the information on their websites and users relatively sure that third parties will not scrape their data, so they can continue to use the websites of their choice for the purposes they intend.

## CONCLUSION

The idea that any and all information on the web is openly accessible and available and therefore can be freely crawled and scraped is wrongheaded. This article demonstrates that actors engaged in these practices across sectors should be aware of the legal factors that discourage crawling and scraping websites for large amounts of data, and the ethical and social factors that argue in favor of close control of crawling in some cases.

Clearly establishing and strengthening legal rules and accountability mechanisms that regulate government, the private sector, academia, and individuals is necessary. The CFAA and trespass doctrine may operate to keep any type of actor from crawling a website and gathering information, but the application of those laws to the internet is unclear, and it can be difficult for the crawled, particularly smaller institutions, to protect themselves under those laws. The government may be further bound by the Fourth Amendment, though the judiciary has yet to make it clear how the Third-Party Doctrine and aggregation principle should bear on the Fourth Amendment in the electronic world and on the internet. Even academia is bound by relatively lax rules, governed only by IRBs.

Without stronger rules and greater accountability, internet users are left open to severe privacy invasions. Their blogs, Facebook and Twitter pages, reviews, photos, discussions on forums can all be scraped, saved, analyzed, and used later for purposes and by people that the users never intended. Though many actors have some rules self-governing their use of crawlers, the rules as a whole are too weak, and holding them accountable is too difficult.

This article presented a number of issues that need to be considered when updating the existing rules governing online surveillance using web crawlers. These issues need to be considered

---

https://searchengineland.com/is-this-googles-twitter-integration-into-the-search-results-220240.

in writing these new rules. Failing to consider them could result in laws that continue to protect a too-narrow view of privacy, or that fail to prevent all the harms that could befall internet users.