Vanderbilt Journal of Entertainment & Technology Law

Volume 20
Issue 3 *Issue 3 - Spring 2018*

Article 4

2018

A Free Ride: Data Brokers'Rent-Seeking Behavior and the Future of Data Inequality

Krishnamurty Muralidhar

Laura Palk

A Free Ride: Data Brokers' Rent-Seeking Behavior and the Future of Data Inequality

Laura Palk[*], *Krishnamurty Muralidhar*[**]

ABSTRACT

Historically, researchers obtained data from independent studies and government data. However, as public outcry for privacy regarding the government's maintenance of data has increased, the discretionary release of government data has decreased or become so anonymized that its relevance is limited. Research necessarily requires access to complete and accurate data. As such, researchers are turning to data brokers for the same, and often more, data than they can obtain from the government. Data brokers base their products and services on data gathered from a variety of free public sources and via the government-created Internet. Data brokers then recategorize the existing free data and combine them with privately collected data. They sell the linked data at a profit while simultaneously preventing the public, whose data they sold, from learning how the data were gathered based on their trade secret protections. To the Authors' knowledge, research has not explored data brokers' rent-seeking behavior and how it will further inequality in accessing credible data—or "data inequality." The Authors contend that without a federal mission to ensure cost-free access to personal data for research and public access purposes, data brokers' sale of such data will potentially lead to biased or inaccurate research results. This development would further the interests of the educated wealthy at the expense of the general public. To resolve this growing data inequality, this Article recommends a variety of legal and voluntary solutions.

* Lecturer, Legal Studies and Accreditation and Assurance of Learning Coordinator, University of Oklahoma Price College of Business and Assistant Adjunct Professor, University of Oklahoma College of Law.

** Professor, Marketing & Supply Chain Management, University of Oklahoma Price College of Business.

TABLE OF CONTENTS

I. INTRODUCTION

> "Rent seeking" is one of the most important insights in the last fifty years of economics and, unfortunately, one of the most inappropriately labeled. . . . The idea is simple but powerful. People are said to seek rents when they try to obtain benefits for themselves through the political arena. They typically do so by getting a subsidy for a good they produce or for being in a particular class of people, by getting a tariff on a good they produce, or by getting a special regulation that hampers their competitors.[1]

Leading economists agree that rent seeking is detrimental to a free-market economy and leads to a decline in growth.[2] Rent-seeking behavior exacerbates income inequality, and thus other forms of inequality, by bending the "rules of some system to shuttle more compensation [to the wealthy]."[3] Generally speaking, if a business is not adding value to the economy but is reaping financial rewards regardless, it is rent seeking.[4] To the Authors' knowledge, research has not explored data brokers' rent-seeking behavior and how it will further inequality in accessing credible data—or "data inequality."[5] The Authors contend that data brokers' sale of personal data, without a concomitant federal mission to ensure cost-free access to such data for research and public access purposes, will potentially lead to biased and inaccurate—or at least uncorroborated and unchallenged—research results. This development would further the interests of the educated wealthy at the expense of the general public.

---

1. *See* THE CONCISE ENCYCLOPEDIA OF ECONOMICS (David R. Henderson ed., 2008), *reprinted in* David R. Henderson, *Rent Seeking*, LIBR. ECON. & LIBERTY, http://www.econlib.org/library/Enc/RentSeeking.html [https://perma.cc/AXY6-TUN9] (last visited Feb. 7, 2018). Gordon Tullock originated this idea in 1967, and Anne Krueger introduced the label "rent seeking" in 1974. Henderson, *supra*.

2. *See* Jim Tankersley, *A Big-Shot Venture Capitalist Says We Need Inequality. What Do Economists Say?*, WASH. POST (Jan. 14, 2016), https://www.washingtonpost.com/news/wonk/wp/2016/01/14/what-silicon-valley-doesnt-understand-about-inequality/?utm_term=.8f8413012607 [https://perma.cc/5WAD-ACZN].

3. *Id.*

4. *See id.* As part of income inequality's negative effect on the economy, economists have concluded the rich are enriching themselves at the expense of their workers. *Id.*

5. In *Privacy and the Varieties of Informational Wrongdoing*, *in* READINGS IN CYBERETHICS 488, 493 (Richard A. Spinello & Herman T. Tavani eds., 2004), Dutch theorist Jeroen van den Hoven coined a similar theory of "information inequality" based on the lack of transparency of data brokers' automation and collection efforts, which was further discussed by Nate Cullerton in the context of data collection and credit scoring. *See* Nate Cullerton, Note, *Behavioral Credit Scoring*, 101 GEO. L.J. 807, 819–20 (2013). The Authors expand on this concept, bringing into focus not only the potential discriminatory uses of opaque data collection but also the lack of privacy regulation placed on data brokers. Combined with the government's practice of declining to release public data based on extreme privacy concerns, the lack of privacy regulation on data brokers creates data manipulation and destructive rent-seeking behavior. *See* discussion *infra* Parts II, V, VII.

Today it is arguably easier to purchase detailed data about a population from data brokers than it is to request such data from the government, and the purchased data cannot be further examined or corroborated because of the data brokers' intellectual property protections. Naturally, well-funded researchers or entities in collaboration with data brokers will have more opportunities to publish research than less well-funded researchers or the general public. Although the Authors acknowledge that this has always been the case, a further imbalance in accessing data that cannot be corroborated will lead to a select few in control of a significant majority of research publications, creating negative rent seeking and data inequality. The Authors conclude data brokers must share data with researchers and the government to further the public welfare without trade secret limitations, and the government must be more flexible in disclosing personal data—particularly where individuals have likely already placed that data into data brokers' hands through their use of the Internet. Finally, the Authors recommend that the data brokers encourage transparency for their underlying research through a self-regulatory incentive similar to a "Fair Trade" designation for consumer products.[6] The trademark of "Transparent Data" could be assigned to those entities who willingly allow the underlying methodology of their data to be corroborated and challenged by researchers. In those instances where a data broker wishes to retain the data's secrecy, the data and any research based on such data could include a disclaimer indicating trade secret protection has been asserted—for example, "data utilized or provided herein is protected by the providers' intellectual property rights and is not subject to corroboration."

By examining the source of a data broker's underlying business, rent seeking becomes apparent. Rent seeking is a theory of economic behavior that entails asking the government for certain privileges or deriving significant profits and advantages without adding any value to the economy.[7] More simply, it consists of transferring wealth rather than creating wealth.[8] This behavior is criticized as contributing to economic inefficiency and economic inequality, as the wealthy receive the benefits of anticompetitive rent-seeking behavior while the rest of the market suffers the losses.[9] Rent seeking as an economic theory has

6. *See Our Global Model*, FAIR TRADE CERTIFIED, https://www.fairtradecertified.org/why-fair-trade/our-global-model [https://perma.cc/RLQ3-2EJ7] (last visited Feb. 7, 2018).

7. *See* Mark Seidenfeld & Murat C. Mungan, *Duress as Rent-Seeking*, 99 MINN. L. REV. 1423, 1426 n.18 (2015); Tankersley, *supra* note 2.

8. Seidenfeld & Mungan, *supra* note 7, at 1426 n.18.

9. *See* Joseph P. Tomain, *Gridlock, Lobbying, and Democracy*, 7 WAKE FOREST J.L. & POL'Y 87, 101, 110 (2017).

been discussed for decades and involves special interest and lobbying efforts in the form of tax relief, subsidies, and preferential regulation in favor of big businesses that have a "symbiotic" relationship with the government.[10]

In Part II of this Article, the Authors examine data brokers' rent-seeking behavior and the legal obstacles posed by potential constraints on this behavior. Research necessarily requires access to complete and accurate data.[11] Based on the ease with which data can be purchased, researchers are turning to data brokers for the same, and often more, data than they can obtain from the government.[12] Data brokers engage in negative rent seeking when they obtain free information from the public and the government, then sell the ensuing data at a profit while simultaneously asserting trade secret protections to prevent the public—whose data they sold—from learning how the

10. *See* Todd Zywicki, *Rent-Seeking, Crony Capitalism, and the Crony Constitution*, 23 SUP. CT. ECON. REV. 77, 78–79 (2015) (citing Mancu Olson's 1982 work *The Rise and Decline of Nations*, which addressed how interest groups capitalize on their power and influence over legislators to obtain special favors). A common illustration of illegal rent-seeking behavior's societal costs is the common criminal. *Id.* at 80–81. The criminal forgoes other productive activity, including gainful employment, and diverts third parties' resources, causing them to purchase theft insurance, alarms, etc., rather than engaging in otherwise productive endeavors and purchases. *Id.* Another common example of negative rent seeking is the rate at which capital gains taxes are calculated. *See, e.g.*, Joseph E. Stiglitz, Opinion, *A Tax System Stacked Against the 99 Percent*, N.Y. TIMES (Apr. 14, 2013, 9:36 PM), https://opinionator.blogs.nytimes.com/2013/04/14/a-tax-system-stacked-against-the-99-percent/?_r=0 [https://perma.cc/6USB-HDP9]. Many of the country's wealthiest individuals are paying taxes only on their carried interest (i.e., their passive investments) rather than on actively earned income because they are not engaged in active employment. *Id.* This is the rent-seeking aspect of their profits, and many legislators have called for reforms that would require the carried interest income be taxed at the individual's ordinary income rate to avoid the consequences of negative rent-seeking behavior. *See id.*; *see also* Michael Cragg & Rand Ghayad, *Inequalities in Tax Policy*, HUFFINGTON POST (May 4, 2015, 7:30 PM), http://www.huffingtonpost.com/rand-ghayad/inequities-in-tax-policy_b_7209108.html [https://perma.cc/TP4S-44ZC].

11. *See generally* Jillian Raines, Note, *The Digital Accountability and Transparency Act of 2011 (Data): Using Open Data Principles to Revamp Spending Transparency Legislation*, 57 N.Y.L. SCH. L. REV. 313, 344 (2012) (discussing the federal legislation designed to inform the public about tracking federal spending).

12. *See* EDITH RAMIREZ ET AL., FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY iv, 3 (2014) [hereinafter 2014 DATA BROKER REPORT], https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf [https://perma.cc/JG3C-ZB9U]; J.H. Reichman & Paul F. Uhlir, *Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment*, 66 LAW & CONTEMP. PROBS. 315, 323 (2003); *see also* Kelsey L. Zottnick, Note, *Secondary Data: A Primary Concern*, 18 VAND. J. ENT. & TECH. L. 193, 200 (2015) (noting that data brokers exploit the laws governing patient confidentiality to sell unprotected information to drug companies).

data were gathered.[13] Next, Part III discusses the public's right to be informed about the internal workings of the government and about the data it maintains along with the current trend toward a "privacy first" philosophy, which constrains governmental officials in their discretionary release of data that should otherwise be considered public.[14] Part IV then explores the advent of "big data" and complications for researchers in accessing largely obscure commercial data compared to open government records. In Part V, the Authors address the lack of regulation in the area of privacy obligations and data brokers. Expanding on this topic in Part VI, the Authors discuss the unclear nature of how data brokers gather and assimilate data in a manner that restricts corroboration and can lead to intentional or unintentional data manipulation, potentially altering the accuracy of research results. Finally, Part VII asserts that fewer privacy protections and more data-sharing incentives should exist between data brokers and the government or general public to avoid the furtherance of "data inequality."

II. DATA BROKERS CONTRIBUTE TO NEGATIVE RENT-SEEKING BEHAVIOR

Data brokers' use of the government-created Internet, combined with profiteering from freely provided information, is a form of negative rent seeking. The US government originally created the Internet as a military tool to collect, store, and decentralize data.[15] Today a nonprofit entity manages the technical aspects of the Internet by assigning connectivity and root management through a public-private contract.[16] Over time, the government allowed commercial entities access to the Internet and abstained from its control.[17] Commercial entities began

13. *See* 2014 DATA BROKER REPORT, *supra* note 12, at 16, 19; *see also* Mary Madden et al., *Privacy, Poverty, and Big Data: A Matrix of Vulnerabilities for Poor Americans*, 95 WASH. U. L. REV. 53, 86 (2017).

14. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1729–30 (2010).

15. *See* Kim Ann Zimmermann & Jesse Emspak, *Internet History Timeline: ARPANET to the World Wide Web*, LIVE SCI. (June 27, 2017, 10:46 AM), www.livescience.com/20727-internet-history.html. In 1969, the US Department of Defense created the precursor to the Internet for military use. *Id.* Additionally, the National Science Foundation maintained control of the Internet hardware. *Id.* Over time, these governmental agencies outsourced their obligations, and the federal government began allowing private commercial entities access to the Internet, resulting in the creation of Google and the world wide web. *See* Victoria D. Baranetsky, *Social Media and the Internet: A Story of Privatization*, 35 PACE L. REV. 304, 325 (2014).

16. *See* Rolf H. Weber & Shawn Gunnarson, *A Constitutional Solution for Internet Governance*, 14 COLUM. SCI. & TECH. L. REV. 1, 6–8 (2013).

17. *See id.* at 9–10.

utilizing the Internet, and the big data industry is expected to reach $47 billion in profits in 2017.[18] Data brokers have taken over the data field initially created by the government and supplied by its users.[19] Data brokers sell data that users freely input through government-funded technology.[20] With today's technology, there is no meaningful consent to the use of our private data.[21] Users do not understand the extent to which their data can be aggregated and further used to extract their financial expenditures.[22] Users of social media platforms, apps, and shopping sites, and those who register their email addresses, are not adequately advised that entities collect their data and transfer it to third parties that aggregate the individuals' data in order to market goods, services, or political ideas back to the individuals based on the data they freely provided.[23] Accordingly, data brokers' profits are based in significant part on technology and information created by tax dollars and the general public without a concomitant privacy or disclosure obligation.

Similar to legal scholars who have likened rent seeking to contractual duress,[24] the Authors contend that data brokers' reuse of a person's aggregated data without that person's clear affirmative consent is a form of inappropriate rent seeking, leading ultimately to further economic and data inequality. Where the contracting party forces another party to consent to contract where he might not otherwise have done so via economic or other duress, the threat-maker can be said to have engaged in detrimental rent-seeking behavior through the transfer of wealth without coordinate value provided by the threat-maker.[25] At its core, rent seeking can be seen as the

---

18. *See* Linda K. Breggin & Judith Amsalem, *Big Data and the Environment: A Survey of Initiatives and Observations Moving Forward*, 44 ENVTL. L. REP. 10984, 10986 (2014).

19. *See id.* at 10985.

20. The US Department of Commerce predicted that private sector profit from government data ranges from $24 billion to $221 billion per year. *See* Frederik Zuiderveen Borgesius et al., *Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework*, 30 BERKELEY TECH. L.J. 2073, 2080–82 (2015).

21. *Id.*; *see* Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1880, 1886 (2013) (describing consumers' lack of knowledge regarding the amount of their data that is used by private data brokers).

22. *See* Solove, *supra* note 21, at 1886.

23. *See id.*

24. *See, e.g.*, Seidenfeld & Mungan, *supra* note 7, at 1437. However, there is a countervailing argument that data mining adds value by collecting, sifting, consolidating, and distributing data in new ways that would otherwise be prohibitively time consuming. *See* Michael Mattioli, *Disclosing Big Data*, 99 MINN. L. REV. 535, 542–44 (2014); Ruth L. Okediji, *Government as Owner of Intellectual Property? Considerations for Public Welfare in the Era of Big Data*, 18 VAND. J. ENT. & TECH. L. 331, 335 (2016).

25. Seidenfeld & Mungan, *supra* note 7, at 1437.

commercialization of existing resources without the input of additional value, as opposed to profit seeking that embodies mutually beneficial transactions.[26] Where efforts are rewarded in the form of wealth redistribution from the rent-seeking behavior rather than by earning wealth through productive activity, the rent-seeking behavior will increase, thus reducing productive jobs and creativity from society.[27] For example, technology industry participants have utilized various forms of rent-seeking behavior to gain regulatory advantages for their fields, as well as to drive their competitors out of business. Recently, Microsoft lobbied the Federal Trade Commission (FTC) to investigate Google, its competitor, for antitrust violations, which cost Google approximately $25 million in counterlobbying efforts.[28] Ultimately, Google succeeded in averting an antitrust suit, which some have hypothesized was due to its frequent access to the Obama administration and key government decision-makers.[29] In this regard, rent-seeking behavior and successful lobbying—rather than a successful market venture—dictate an entity's economic survival.[30]

Data brokers, and some scholars, will likely contend that they are not rent seekers, as they provide added value by aggregating discrete data sets through independently created algorithms, thus allowing third parties to develop a fuller picture about consumers and providing consumers with pertinent information.[31] Because these algorithms are the data brokers' protected trade secrets, the general

---

26. Big data has special value and "resides far downstream from the commercial exchanges that take place between data producers and their customers." *See* Mattioli, *supra* note 24, at 549.

27. *See* Zywicki, *supra* note 10, at 83.

28. *See id.* at 84.

29. *See* Johnny Kampis, *Visitor Logs Show Google's Unrivaled White House Access*, WATCHDOG (May 16, 2016), https://www.watchdog.org/issues/accountability/visitor-logs-show-google-s-unrivaled-white-house-access/article_1b3bf08d-1776-5ee1-8287-0df696f8ec37.html [https://perma.cc/RZ22-HGYD]; Brody Mullins, *Google Makes Most of Close Ties to White House*, WALL ST. J. (Mar. 24, 2015, 9:24 PM), https://www.wsj.com/articles/google-makes-most-of-close-ties-to-white-house-1427242076 [https://perma.cc/ZP59-FT9D].

30. *See* Zywicki, *supra* note 10, at 102–03. Although lobbying and other forms of political rent seeking are legal, as opposed to forms of illegal rent seeking such as outright theft, the economic effects are similar. *Id.* at 80–82. The opportunity costs associated with seeking governmental privileges include travel costs for corporate officers to meet with politicians, campaign contributions, and money spent influencing officials rather than corporate officers utilizing their time to manage their business in a more efficient manner. *Id.* at 81.

31. *See* Okediji, *supra* note 24, at 334; *What Are Data Brokers—And What Is Your Data Worth?*, WEBPAGEFX (Apr. 16, 2015), https://www.webpagefx.com/blog/general/what-are-data-brokers-and-what-is-your-data-worth-infographic/ [https://perma.cc/WLU4-DCRQ] (detailing how data brokers often obtain data and how much some users are paid for their data, noting the majority of data comes from public data or consumers' voluntary input of data into a variety of sources like loyalty programs).

public does not have access to them despite its significant contribution to the data brokers' profit.[32] The aggregated data are then sold to third parties for marketing purposes.[33] In this regard, data brokers simply access free data from individuals who are generally unaware that their data are being repurposed and sold to third parties who wish to market to the same individuals and obtain further purchases from them.[34] Although entirely legal, the combination of these three factors—(1) individuals freely providing data without realizing the data's resale value, (2) legally protected aggregation of individuals' data, and (3) use of this data to convince the users to purchase products they might not have otherwise bought—arguably represents the transfer of wealth rather than the creation of wealth.[35] The Authors contend that this is a form of detrimental rent seeking in need of reform.

Despite the negative consequences to the economy from this rent-seeking behavior, the US Constitution is an obstacle to its constraint. Data brokers' aggregation and resale of data is commercial speech protected by the First Amendment.[36] In *Citizens United v. Federal Election Commission*, the US Supreme Court held that corporations are people in the eyes of the First Amendment and have the right to support a political viewpoint and candidate through financial means—rendering legislation on lobbying efforts difficult,

---

32. *See* Madden et al., *supra* note 13, at 86; Ashley Kuempel, Comment, *The Invisible Middlemen: A Critique and Call for Reform of the Data Broker Industry*, 36 NW. J. INT'L L. & BUS. 207, 210 (2016); *What Are Data Brokers—And What Is Your Data Worth?*, *supra* note 31; *see also* Adam M. Samaha, *Government Secrets, Constitutional Law, and Platforms for Judicial Intervention*, 53 UCLA L. REV. 909, 922 (2006) ("Openness exposes not just waste, fraud, and abuse, but also . . . candid advice, intimately private information, and trade secrets. . . . A rule of full disclosure might also prompt officials to sanitize the public record as it is created.").

33. *See* Vivian Adame, *Consumers' Obsession Becoming Retailers' Possession: The Way That Retailers Are Benefiting from Consumers' Presence on Social Media*, 53 SAN DIEGO L. REV. 653, 687 (2016).

34. Scholars may contend that data and marketing materials directed to special interests are likewise added value and not rent seeking. *See* Okediji, *supra* note 24, at 334.

35. More plainly, data brokers use technology which was created at taxpayer expense—i.e., the Internet—to repurpose information which users freely and unwittingly provide to one source. *See* discussion *supra* note 15 and *infra* notes 62–63. Users are unaware that their provided information will be combined with other information they provide to different sources online and then resold. *See* Baranetsky, *supra* note 15, at 337. Users are not provided meaningful disclosure of this aggregation and reuse, nor are they generally compensated. *See* Kuempel, *supra* note 32, at 222. Likewise, data brokers pay nothing to the government (outside of donations and taxes) for their use of the Internet. And while some may argue that the Internet is a free public resource, the sheer amount of profit derived from the Internet and freely provided information warrants some additional consideration, either through disclosure obligations or opt-out mechanisms described throughout this Article.

36. *See* Richard L. Hasen, *Lobbying, Rent-Seeking, and the Constitution*, 64 STAN. L. REV. 191, 198–99 (2012); *see also* Neil M. Richards, *Reconciling Data Privacy and the First Amendment*, 52 UCLA L. REV. 1149, 1176 (2005).

though not impossible.[37] The Court noted that the government could "regulate corporate political speech through disclaimer and disclosure requirements, but it may not suppress that speech altogether."[38] With respect to political rather than commercial speech, the *Citizens United* Court found restrictions on political speech "are 'subject to strict scrutiny,' which requires the Government to prove that the restriction 'furthers a compelling interest and is narrowly tailored to achieve that interest.'"[39] Regarding the underlying disclaimer and disclosure requirements of the challenged law, the Court determined these provisions are governed under "exacting scrutiny," which requires a "substantial relation" between the disclosure requirement and a "sufficiently important" governmental interest.[40] The Supreme Court found that disclosure requirements are a "less restrictive alternative to more comprehensive regulations of speech."[41] In *Citizens United*, the Court found no evidence that the disclosure requirements as applied would expose the individuals identified by the disclosure to harassment or abuse and determined that they were therefore constitutional.[42] The Court rationalized that the disclosure obligations allow the electorate to be fully informed and give proper weight to the speakers and their messages.[43] Because the essence of the First Amendment is to ensure free and open discourse, the disclosure obligations further an important interest and were found constitutional.[44]

Along these lines, legal scholars have argued that national economic welfare and the idea of income or social inequality are compelling interests warranting narrowly tailored regulations of commercial or political speech.[45] Ultimately, reducing rent-seeking

---

37. Hasen, *supra* note 36, at 195–96; *see* Citizens United v. Fed. Election Comm'n, 558 U.S. 310, 342 (2010).

38. *See Citizens United*, 558 U.S. at 318.

39. *Id.* at 340 (quoting Fed. Election Comm'n v. Wis. Right to Life, Inc., 551 U.S. 449, 464 (2007) (opinion of Roberts, C.J.)).

40. *Id.* at 366; *see id.* at 315 (finding that although disclaimer and disclosure requirements can "burden the ability to speak . . . they do not prevent anyone from speaking." (internal quotations omitted)).

41. *Id.* at 369.

42. *Id.* at 370. There may be times when a disclosure obligation as applied is unconstitutional, such as when there is a "reasonable probability that the group's members would face threats, harassment, or reprisals if their names were disclosed." *Id.*

43. *Id.* at 371.

44. *Id.* However, the Court did find limitations on corporate lobbying and political contributions unconstitutional. *Id.* at 356.

45. *See* Hasen, *supra* note 36, at 198–99. The level of judicial scrutiny depends on the type of behavior being regulated. For example, campaign contributions receive a lower form of First Amendment protection because they are a form of commercial speech, as opposed to bans on political speech, like solicitation prohibitions and revolving door statutes, which are subject to strict scrutiny. *Id.* at 239.

behavior improves economic productivity and can lead to a decrease in the federal budget deficit, while permitting rent-seeking behavior leads to slow economic growth in the long term.[46] Reducing unproductive and anticompetitive behavior that results in wealth inequality is arguably a compelling governmental interest that would withstand even the most exacting judicial scrutiny required by *Citizens United*.[47] Accordingly, specific legislation designed to reduce data brokers' rent-seeking behavior through disclosure obligations should be constitutional. The legislation would require data brokers to advise users of how their data will be used, aggregated, and resold. Further, any research results based on data purchased or affiliated with data brokers should likewise disclose whether third parties may corroborate the underlying research data. More restrictive obligations requiring data brokers to actually provide and share their underlying data with the government or researchers without charge—so that the data could be used, corroborated, and challenged—likely would be unconstitutional unless strictly circumscribed to very limited situations, such as information related to significant public welfare or national security issues.[48]

The inability to curb data brokers' anticompetitive behavior based on First Amendment grounds could be a significant hurdle[49] that will force legislators to address regulations in terms of national economic welfare and inequality. A prime example of a defeated attempt to constrain commercial data brokers' speech is the Supreme Court's ruling in *Sorrell v. IMS Health, Inc.*[50] IMS Health, Inc. (IMS) is a data broker in the healthcare field that analyzes trends for certain companies which, in turn, market to their customers.[51] One service IMS

46. *Id.* at 232 (discussing economists' studies on economic market growth in eras with high and low rent-seeking activity and noting a significant correlation with negative results in high rent-seeking eras).

47. *See* Tomain, *supra* note 9, at 110, 113; *see also Citizens United*, 558 U.S. at 312.

48. An outright requirement of free provision of information could be construed as an unconstitutional taking or conversion of data brokers' property. U.S. CONST. amend. V; *see* Mark A. Lemley, *Private Property*, 52 STAN. L. REV. 1545, 1549 (2000) (stating that allowing individuals to have a property right in the data they contribute online would have a negative impact on commerce). *See generally* Tahoe-Sierra Pres. Council, Inc. v. Tahoe Reg'l Planning Agency, 535 U.S. 302 (2002); Trans Union LLC v. Credit Research, Inc., No. 00 C 3885, 2001 WL 648953 (N.D. Ill. June 4, 2001) (denying a motion to dismiss possible conversion claims where a licensee allegedly exceeded its use agreement for online data); Mark Bartholomew, *Intellectual Property's Lessons for Information Privacy*, 92 NEB. L. REV. 746, 756, 786 (2014) (discussing the trend that data collectors have nearly an undisturbed right to free speech and proposing a balancing analysis for courts).

49. *See* Tomain, *supra* note 9, at 113.

50. Sorrell v. IMS Health Inc., 564 U.S. 552 (2011).

51. *Id.* at 558.

provides is the service of "detailing," by which IMS collects prescription and purchasing data from individual pharmacies, identifies physician trends in prescribing pharmaceuticals, and then profiles the physicians in an effort to assist pharmaceutical companies in marketing their drugs to those doctors.[52] Vermont sought to lower the cost of prescription drugs by restricting data detailing because the costs of such detailing were passed on to consumers.[53] The law prohibited pharmaceutical and insurance companies from selling prescription data to the data brokers.[54] IMS challenged the law on corporate free speech grounds, and the Supreme Court ultimately ruled the law unconstitutional.[55] The Court found the statute was a content-based restriction on the marketer and on the data's use rather than a ban on other forms of speech on the same topic.[56] The Court went on to note that "the creation and dissemination of information are speech within the meaning of the First Amendment."[57]

Despite the *Sorrell* Court's proclamation that disclosure of consumers' identities is protected speech, courts have since determined that legislative regulation of speech may still be appropriate.[58] Speech that is commercial in nature is subject to a lesser standard than strict scrutiny, including speech marketing goods and services.[59] As a result,

---

52. *Id.*

53. *Id.* at 572.

54. *Id.* at 563.

55. *Id.* at 580.

56. *Id.* at 565.

57. *Id.* at 570.

58. *See, e.g.*, Boelter v. Hearst Commc'ns, Inc., 192 F. Supp. 3d 427, 444 (S.D.N.Y. 2016). Lower courts have struggled with the implications of *Sorrell* in the context of compelled disclosures surrounding commercial speech. *See* Note, *Repackaging* Zauderer, 130 HARV. L. REV. 972, 978–83 (2017). The US Court of Appeals for the Sixth Circuit requires the lesser rational basis scrutiny for governmental restrictions on commercial speech that is "likely" to mislead, whereas the Fifth and Eighth Circuits allow a rational basis test where the commercial speech is "potentially misleading." *Id.* at 980; *see also* Zauderer v. Office of Disciplinary Counsel, 471 U.S. 626, 651 (1985) (holding mandatory disclosure obligation of "purely factual and uncontroversial information" aimed at preventing consumer deception does not violate the First Amendment).

59. *See* Central Hudson Gas & Elec. Corp. v. Pub. Serv. Comm'n, 447 U.S. 557, 566 (1980); *Boelter*, 192 F. Supp. 3d at 447. One area where data brokers profit from rent-seeking behavior is in the aggregation of existing medical data that individuals willingly provide to healthcare professionals, to investigative studies, or to other online services. *See* Barbara J. Evans, *Power to the People: Data Citizens in the Age of Precision Medicine*, 19 VAND. J. ENT. & TECH. L. 243, 248 (2016); Adam Tanner, *How Data Brokers Make Money Off Your Medical Records*, SCIENTIFIC AMERICAN (Feb. 1, 2016), https://www.scientificamerican.com/article/how-data-brokers-make-money-off-your-medical-records/ [https://perma.cc/6WT2-6DG4]. With current technology, there is a greater risk that a person's anonymized data can be reverse engineered to reveal his identity. *See* Tanner, *supra*. The most significant data broker in this market is IMS, which recorded $2.6 billion in revenue in 2014. *Id.* Pfizer, the pharmaceutical giant, pays $12 million annually to purchase this type of data from data brokers, including IMS. *Id.*

legal scholars have advocated for regulation in the area of data privacy to include providing consumers with the right to affirmatively consent to and correct the data gathered about them, as well as better privacy protections on the types of data that a data broker may maintain and sell.[60] Whether such regulation would violate *Sorrell* and the First Amendment is the subject of debate.[61] Privacy laws generally do not protect individuals from data brokers' resale of their data.[62] Rather, data brokers and companies rely on both the consumers' acceptance of their terms of use and their broad privacy policies to reuse and repurpose consumer data.[63]

---

60. *See, e.g.*, Borgesius et al., *supra* note 20, at 2128; Kimberly A. Houser & Debra Sanders, *The Use of Big Data Analytics by the IRS: Efficient Solutions or the End of Privacy as We Know It?*, 19 VAND. J. ENT. & TECH. L. 817, 839–41, 871–72 (2017). The US Department of Health, Education and Welfare has developed privacy guidance documents. *See* Borgesius et al., *supra* note 20, at 2109. One of the guidelines is to restrict repurposing data collected for other reasons without consent. *Id.*

61. *See* Neil M. Richards, *Why Data Privacy Law Is (Mostly) Constitutional*, 56 WM. & MARY L. REV. 1501, 1521 (2015).

62. *See* discussion *infra* Part V. Data is collected through a person's browsing history, through a person's purchases, and by tracking their cookies. *See* EDITH RAMIREZ ET AL., FED. TRADE COMM'N, BIG DATA: A TOOL FOR INCLUSION OR EXCLUSION? UNDERSTANDING THE ISSUES 3–4 (2016), https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf [https://perma.cc/732T-4SZU] (detailing how data is gathered, analyzed, and used). Although the Health Insurance Portability and Accountability Act (HIPAA) protects individuals' private medical information, it is inapplicable to data brokers' marketing and research activities. *See* Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (1996) (codified as amended in scattered sections of 18, 26, 29, and 42 U.S.C.); 2014 DATA BROKER REPORT, *supra* note 12, at 14 n.41. Nonetheless, such information can be used or disclosed when certain anonymization techniques eliminate the ability to identify an individual, such as generalizing birth dates, zip codes, etc. *See* Health Insurance Portability and Accountability Act of 1996, § 262, 110 Stat. at 2029–30 (codified as amended at 42 U.S.C. §§ 1320d-6 to d-7 (2012)); 2014 DATA BROKER REPORT, *supra* note 12, at 12. Data brokers are generally exempt from HIPAA obligations because data brokers receive or purchase aggregated and de-identified data from a covered entity, meaning the individual's identity is not provided to the data broker. *See* 2014 DATA BROKER REPORT, *supra* note 12, at 9–10; *see also Covered Entities and Business Associates*, U.S. DEP'T HEALTH & HUM. SERVS., http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/ [https://perma.cc/39DU-AW94] (last visited Feb. 8, 2018); *To Whom Does the Privacy Rule Apply and Whom Will It Affect?*, U.S. DEP'T HEALTH & HUM. SERVS., https://privacyruleandresearch.nih.gov/pr_06.asp [http://perma.cc/N3FJ-57P7] (last visited Feb. 8, 2018). The data brokers can aggregate the data along with statistical, de-identified data gathered from pharmacies and insurance companies, creating a valuable data commodity that third parties purchase to educate them on future investments and marketing schemes. *To Whom Does the Privacy Rule Apply and Whom Will It Affect?*, *supra*. Although personally identifiable health data are technically confidential and only statistical data about individuals are gathered and aggregated, the reality is that a third party can discover a person's identity from this data. *See id.*

63. *See* Solove, *supra* note 21, at 1886 (describing consumers' lack of knowledge regarding the amount of their data that is used by private data brokers).

With these two Supreme Court cases as a backdrop, any governmental restrictions on data brokers' powers to collect and sell data must withstand significant judicial scrutiny. The manner in which regulations could withstand this scrutiny lies within the realm of disclosure requirements rather than restriction requirements.[64] It is within this area that the Authors contend the lack of regulation will lead to data inequality and arguably violate the public's right to access data, in turn violating the public's First Amendment rights.[65]

III. MODERN COMPLICATIONS FOR A RESEARCHER'S ACCESS TO DATA

Data brokers control the nature of and access to data that will form the basis of future research. Many data brokers and other technology companies collaborate with researchers to conduct a variety of research, but the underlying data and processes surrounding their research are never made public.[66] In this regard, the underlying data cannot be checked or challenged.[67] If access to similar data is unavailable without purchase from the data broker, then it follows that only the wealthiest researchers or those with special relationships with data brokers will have their voices heard, detrimentally impacting the national public welfare. Accordingly, data brokers should be required to share certain big data necessary for public research with the government—a form of disclosure requirement subject to the Supreme Court analysis noted above.

Big data has altered human subject research, including expanding the definition of who is a researcher.[68] Innovative application developers and others collect significant amounts of data without oversight.[69] These flexible practices and the public's inability

64. *See* Richards, *supra* note 61, at 1521.

65. *See* Richmond Newspapers, Inc. v. Virginia, 448 U.S. 555, 583 (1980) (holding that where the government creates an arbitrary obstacle to important data, it is violating the First Amendment rights of the person attempting to access the data).

66. *See, e.g.*, Dustin Volz, *Facebook Inks Agreement with 17 Universities to Streamline Research*, REUTERS (Dec. 21, 2016, 1:27 PM), https://www.reuters.com/article/us-facebook-research/facebook-inks-agreement-with-17-universities-to-streamline-research-idUSKBN14A2AX_[https://perma.cc/W7FN-83TR]; Jaikumar Vijayan, *Google Invites University Researchers to Collaborate on IoT Projects*, EWEEK (Feb. 12, 2016), http://www.eweek.com/networking/google-invites-university-researchers-to-collaborate-on-iot-projects [https://perma.cc/BU74-WHRJ].

67. *See, e.g.*, Reichman & Uhlir, *supra* note 12, at 319–320, 354 (analyzing the nature of scientific data, the advent of data brokers, and their desire to protect their research outcomes through intellectual property mechanisms); *id.* at 427 (recommending additional university and governmental public research).

68. *See* Omer Tene & Jules Polonetsky, *Judged by the Tin Man: Individual Rights in the Age of Big Data*, 11 J. TELECOMM. & HIGH TECH. L. 351, 353 (2013).

69. *Id.*

to test the researchers' analysis of private data can lead to a "credibility crisis in computational science, not only among scientists, but as a basis for policy decisions and in the public mind."[70] Scholars submit that the lack of transparency regarding what data are used, how data are collected, and how data are analyzed are significant issues with new research.[71] In this regard, access to public data is paramount to the future of credible and unbiased research. Notably, the public obtains data from two primary sources: (1) private data miners (e.g., data brokers, social media, search engines, app providers), and (2) the US government.[72] The US government provides data to the public in three general forms: (1) through the general release of mass data, (2) upon request for records maintained by the government under the Freedom of Information Act (FOIA),[73] and (3) through governmentally funded research.[74]

A. What Is Data?

Before detailing how the government delivers data to the public, it is necessary to explain what this Article means by data. Data is defined as the "representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process."[75] Datafication is the "act of rendering these representations into a format that can be communicated or manipulated by some process."[76] Researchers often utilize a combination of sources, including public data and commercial data.[77] Further, agencies release data through a variety of means, including "derived index data, aggregated tables or sanitized microdata in public use data files, raw data controlled via a

---

70. *Id.* at 354.

71. *See, e.g., id.* at 355.

72. *See generally* Jennifer Bresnahan, *Personalization, Privacy, and the First Amendment: A Look at the Law and Policy Behind Electronic Databases*, 5 VA. J.L. & TECH. 8 (2000) (discussing the strong constitutional protections for data brokers, their databases, and their data mining practices).

73. Freedom of Information Act, Pub. L. No. 90-23, 81 Stat. 54 (1967) (codified as amended at 5 U.S.C. § 552 (2012)).

74. *See* Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L.J. 1967, 1991 (2015); Reichman & Uhlir, *supra* note 12, at 396.

75. *See* Meg Leta Ambrose, *Lessons from the Avalanche of Numbers: Big Data in Historical Perspective*, 11 I/S: J.L. & POL'Y FOR INFO. SOC'Y 201, 210 (2015) (quotations omitted).

76. *Id.* at 211 (emphasis omitted).

77. *See* Altman et al., *supra* note 74, at 2001. Unlike public data sets, restricted data requires researchers to apply for access to the data, and the governmental release depends on a formal screening process. *Id.* at 1996. The use is limited to the purposes specified through data use agreements. *Id.* at 1996.

secure data enclave, [and], to a lesser extent, data made available online through query systems."[78] The term "big data" includes the "novel ways in which organizations including government and businesses[] combine diverse digital datasets and then use statistics and other datamining techniques to extract from them both hidden and surprising correlation[s]."[79] Big data begins in the form of small segments of data collected, consolidated, and analyzed.[80] Entities like advertising networks, social media, banks, and retailers analyze the data and build consumer profiles, store billions of data elements on consumers, and then predict how the consumer will behave based on the profile.[81] Data mining, meanwhile, is the complex process of taking data collected from a variety of sources—both public and private—removing unreliable or redundant data, and constructing statistical models using the remaining data such that anyone in possession of the mined data can predict future behaviors.[82] It is this aggregation that data brokers wish to protect and would assert adds value to the economy.[83] However, the added value is at the extreme expense of unwitting users and the research community, with inordinate profit to the data brokers who would not exist without the government-created Internet and without users providing them with free information.[84]

Every second, individuals in the United States tweet approximately six thousand times, enter forty thousand Google searches, and send over two million emails.[85] By the year 2014, the Internet contained one billion websites.[86] There are generally three types of relevant research data: (1) aggregated data (summary information released to the public),[87] (2) de-identified microdata released to researchers for analytical purposes ("data [released] in its

78. *Id.* at 1993.

79. *See* Ambrose, *supra* note 75, at 212 (quoting Ira S. Rubinstein, *Big Data: The End of Privacy or a New Beginning?*, 3 INT'L DATA PRIVACY L. 74, 74 (2013)). Big data "refers to a new method of empirical inquiry." Mattioli, *supra* note 24, at 539.

80. *See* Mattioli, *supra* note 24, at 539.

81. Courtney A. Barclay, *Protecting Consumers by Tracking Advertisers Under the National Broadband Plan*, 19 MEDIA L. & POL'Y 57, 67 (2010).

82. *See* Zottnick, *supra* note 12, at 196.

83. *See* Reichman & Uhlir, *supra* note 12, at 354, 368–69.

84. *Id.* at 371.

85. *See* Stephanie Pappas, *How Big Is the Internet, Really?*, LIVE SCI. (Mar. 18, 2016, 11:40 AM), www.livescience.com/54094-how-big-is-the-internet.html [https://perma.cc/7YLN-TSUS].

86. *Id.*

87. *See* ROBERT I. KABACOFF, R IN ACTION: DATA ANALYSIS AND GRAPHICS WITH R 112 (2011).

most granular, unaggregated form"),[88] and (3) identified data (customer identification, at least in some form like an IP address, necessary for targeted marketing purposes).[89] The importance of big data and a data broker's role in today's research cannot be overstated.

B. The Importance of Publicly Available Data

Historically, the general public and academic researchers relied on data gathered and disseminated by "public institutions"—including government agencies, nonprofit organizations, universities, and research centers—by accessing routinely publicized agency data releases.[90] Public data, or the "data commons," inform the public and enhance research.[91] Increased demand for privacy in recent years has led government agencies to be less inclined to share their data or more inclined to enact data protection measures that diminish the utility of the data.[92] Simultaneously, there has been an exponential increase in the quantity and quality of data collected by private sources.[93] This collected data can be purchased with or without disclosing individual identities.[94]

The increase in demand for privacy is arguably driven primarily by private sources collecting and selling the data, yet there are minimal,

---

88. MATTHEW RUMSEY, CTR. FOR OPEN DATA ENTER., BRIEFING PAPER ON OPEN DATA AND PRIVACY 2 (2016), http://reports.opendataenterprise.org/BriefingPaperonOpenDataandPrivacy.pdf [https://perma.cc/CDS9-BZW3].

89. *See* Marcia M. Boumil et al., *Prescription Data Mining, Medical Privacy and the First Amendment: The U.S. Supreme Court in* Sorrell v. IMS Health Inc., 21 ANNALS HEALTH L. 447, 450 (2012); Ira S. Rubinstein, *Voter Privacy in the Age of Big Data*, 2014 WIS. L. REV. 861, 924–25 (2014).

90. *See* Reichman & Uhlir, *supra* note 12, at 331–33.

91. *See* Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. TECH. 1, 2–3 (2011). Reuse of public data creates new business, services, and productivity. *See* Farnam Jahanian, *The Policy Infrastructure for Big Data: From Data to Knowledge to Action*, 10 I/S: J.L. & POL'Y FOR INFO. SOC'Y 865, 866–68 (2015). For example, financial service providers use statistics for input, and the meteorological field uses weather data to provide specific forecasting for offshore oil companies. *See* Borgesius et al., *supra* note 20, at 2081.

92. *See* J. Trent Alexander, Michael Davern & Betsey Stevenson, *Inaccurate Age and Sex Data in the Census PUMS Files: Evidence and Implications* 1–3 (CESifo, Working Paper No. 2929, 2010), http://ssrn.com/abstract=1546969 [https://perma.cc/8UFR-FKE7]; *Can You Trust Census Data?*, FREAKONOMICS (Feb. 2, 2010, 11:09 AM), http://freakonomics.com/2010/02/02/can-you-trust-census-data/ [https://perma.cc/SH7L-WMMU]; *see also* Lara Cleveland et al., *When Excessive Perturbation Goes Wrong and Why IPUMS-International Relies Instead on Sampling, Suppression, Swapping, and Other Minimally Harmful Methods to Protect Privacy of Census Microdata, in* PRIVACY IN STATISTICAL DATABASES 179, 181 (Josep Domingo-Ferrer & Ilenia Tinnirello eds., 2012).

93. *See* Joseph A. Tomain, *Online Privacy & the First Amendment: An Opt-In Approach to Data Processing*, 83 U. CIN. L. REV. 1, 3 (2014).

94. *Id.*

if any, private data aggregator privacy requirements (legal or otherwise).[95] In most cases, the private sources rely on the user's "consent," usually in the form of clicking a box when using a website or downloading an application.[96] The very availability of data from these private sources has led some to demand that the *government* enhance its data protection to prevent those with access to the private sources of data from potentially reverse engineering the individual's identity before accessing the government's records.[97] To the Authors, this seems unfair, forcing the government to protect individual privacy while private data brokers bear limited similar burdens.

IV. OPEN ACCESS TO DATA WITHIN THE GOVERNMENT'S CONTROL

There has been considerable research on the practices of government agencies both from a technical and a policy perspective.[98] But to the Authors' knowledge, the effect of increased governmental privacy obligations on researchers and the public has not been examined. Central to a democratic environment is a philosophy of the citizenry's right to know about the internal workings and decisions of its government and the data that it maintains.[99] Although American colonists believed in the idea of the public's right to know, the US Constitution does not contain such a provision; in fact, the Founding Fathers were less than transparent in their management of the government.[100] It was not until 1943, in *Martin v. City of Struthers*, that the Supreme Court first recognized "a constitutional right to *receive* information" under the First Amendment.[101] Thereafter, many states enacted legislation governing the retention and maintenance of

---

95. For examples of such requirements, see 15 U.S.C. §§ 1681, 6501–06, 6801 (2012); 18 U.S.C §§ 2710, 2721–25 (2012); Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (1996) (codified as amended in scattered sections of 18, 26, 29, and 42 U.S.C. (2012)). *See also* Paul M. Schwartz, *Privacy and Democracy in Cyberspace*, 52 VAND. L. REV. 1609, 1611 (1999) (noting the lack of standards for cyberspace privacy, "legal or otherwise").

96. *See* Tomain, *supra* note 93, at 35–36.

97. *See* Nancy S. Kim & D.A. Jeremy Telman, *Internet Giants as Quasi-Governmental Actors and the Limits of Contractual Consent*, 80 MO. L. REV. 723, 728–29 (2015); Anne Klinefelter, *When to Research Is to Reveal: The Growing Threat to Attorney and Client Confidentiality from Online Tracking*, 16 VA. J.L. & TECH. 1, 40–41 (2011).

98. *See* Andrew Chin & Anne Klinefelter, *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study*, 90 N.C. L. REV. 1417, 1427 (2012).

99. *See* David Cuillier, *The People's Right to Know: Comparing Harold L. Cross' Pre-FOIA World to Post-FOIA Today*, 21 COMM. L. & POL'Y 433, 438 (2016). This concept dates back to the Athenians in 330 BC. *Id.*

100. *Id.* at 439 (explaining how the early stages of the US government acted in secrecy).

101. *Id.* (emphasis in original); *see* Martin v. City of Struthers, 319 U.S. 141, 143 (1943) (holding that freedom of speech protections encompass the "right to distribute literature . . . and necessarily protects the right to receive it").

government records, and the federal government enacted the Administrative Procedure Act,[102] establishing internal operating and records retention procedures for federal agencies.[103] Many secrecy and censorship laws originated with World War II for national security reasons.[104] After the war, President Harry S. Truman continued classifying many records as "secret," drawing significant and widespread journalistic criticism.[105] These concerns led to the American Society of News Editors' report in 1953, addressing "customs, laws and court decisions affecting our free access to public information whether it is recorded on police blotters or in the files of the national government."[106] FOIA was enacted as a result of this report and is one form of governmental release of records.[107]

Outside the context of FOIA, a second method of government data release is common public-sector data publication, including government performance data.[108] Government performance data are defined as data that "can be freely used, modified, and shared by anyone for any purpose."[109] Governments struggle with the benefits of releasing public data and the potential privacy implications.[110] However, several federal agencies routinely release public sector data.[111] For example, the Census Bureau releases statistical data about individuals in an aggregated form gathered from interviews and questionnaires, creating official statistics from tabular or relational data.[112] Voluntary release of data provides for transparent research—as opposed to research protected by intellectual property laws—and enables other researchers to test the original researcher's

102. Administrative Procedure Act, Pub. L. No. 79-404, 60 Stat. 237 (1946) (codified as amended at 5 U.S.C. §§ 551–59 (2012)).

103. *See* Cuillier, *supra* note 99, at 441.

104. *Id.* at 440.

105. *Id.*

106. *Id.* at 441 (quoting HAROLD L. CROSS, THE PEOPLE'S RIGHT TO KNOW: LEGAL ACCESS TO PUBLIC RECORDS AND PROCEEDINGS xv (1953)).

107. *See id.* at 442–43 (providing a detailed history of the basis for FOIA, which took many of its provisions from excerpts of state laws, common law, case law, attorney general opinions, and agency regulations as half of the states had public disclosure laws).

108. *See, e.g.*, *Census Data Mapper*, U.S. CENSUS BUREAU, https://www.census.gov/geo/maps-data/maps/datamapper.html [https://perma.cc/35RT-SAUS] (last visited Feb. 8, 2018).

109. *See* Borgesius et al., *supra* note 20, at 2076 (quotations omitted).

110. *Id.*

111. *See, e.g.*, *Census Data Mapper*, *supra* note 108.

112. *See* Altman et al., *supra* note 74, at 1991. Certain government agencies gather and release statistical data to assist government policy and economic decisions, research, and transparency. *Id.*

analysis and opinion for a more thorough examination of the topic.[113] Despite the importance of publicly available data, individual privacy in the data is likewise significant.

A. Open Access to Governmental Data Under FOIA Versus Personal Privacy

While access to information is important for researchers, individual privacy interests in protecting sensitive data are also important. One aspect of the "privacy first" analysis is determining when, or even if, the government will release relatively obscure data that involves examining how privacy demands affect data release associated with FOIA requests.[114] The relationship between the researcher and the government agency is very different in the context of the voluntary release of data as opposed to FOIA releases. Government agencies voluntarily release data under legal mandates (as in the case of the Census Bureau) or as an integral part of their function.[115] However, FOIA requests often place the researcher and the agency in an adversarial position because the agency is reluctant to release the data, but the FOIA request forces the data's release.[116] FOIA's main goal is to ensure the public is informed about the government so that it can be held accountable for its actions—a form of "transparency first" analysis.[117]

Despite the premise of transparency first, there are nine FOIA exemptions allowing the government to refuse to release records in the government's possession—in particular, Exemptions 6 and 7 regarding personnel and personal privacy records, respectively.[118] Initially, FOIA exemptions were not designed as "mandatory bars to disclosure."[119] Rather, the exemptions provided the agency with *discretion* to withhold

---

113. *See* Rebecca Lipman, *Online Privacy and the Invisible Market for Our Data*, 120 PENN ST. L. REV. 777, 789–92 (2016).

114. Although FOIA and open records laws often speak in terms of FOIA disclosures, research in this field tends to define a disclosure as an unintentional release of sensitive data rather than the voluntary release of data. *See generally* Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1118–20 (2013). Thus, the Authors utilize the term "release" when data is voluntarily and properly released, as opposed to a "disclosure" relating to the release of data that may contain information leading to de-identification of an individual.

115. *See Census Data Mapper*, *supra* note 108.

116. *See* 5 U.S.C. § 552 (2012).

117. *See* John Doe Agency v. John Doe Corp., 493 U.S. 146, 152 (1989). FOIA specifically provides that the government shall release records "upon any request for records which (i) reasonably describes such records and (ii) is made in accordance with published rules[.]" 5 U.S.C. § 552(a)(3)(A).

118. 5 U.S.C. § 552(b).

119. *See* Chrysler Corp. v. Brown, 441 U.S. 281, 293 (1979).

data in its possession.[120] The exemptions are written in permissive terms, so the government *may* withhold the release of records instead of "prohibiting" their release.[121] Where data do not fall within one of these exemptions, discretionary government release of the data may be permissible and appropriate.[122] However, certain exemptions may be inappropriate for discretionary government release, including Exemption 6 regarding the potential release of personnel and medical records and Exemption 7(C) regarding records that contain information involving one's personal privacy.[123] The lack of clarity regarding when a discretionary release is appropriate, particularly under Exemptions 6 and 7(C) for privacy reasons, causes agencies to err on the side of privacy protection and nonrelease of the data.[124] Evidence that agencies decline to release data under FOIA is revealed through a comparison of FOIA release outcomes under both the Bush and Obama administrations.[125] As shown in Figure 1 below, the majority of declinations are based on privacy concerns, and under the Obama administration those figures increased despite the executive branch's policy for transparent government.[126]

---

120. *Id.* at 294.

121. *See* 5 U.S.C. § 552(b); *see also* Eamon D., *Analyzing FOIA Statistical Trends from FY2011 to FY2012*, AINS (Aug. 28, 2013), http://ains.com/foiablog/2013/8/28/analyzing-foia-statistical-trends-from-fy2011-to-fy2012.html [https://perma.cc/A7ZE-T5AA].

122. *See* Eamon D., *supra* note 121.

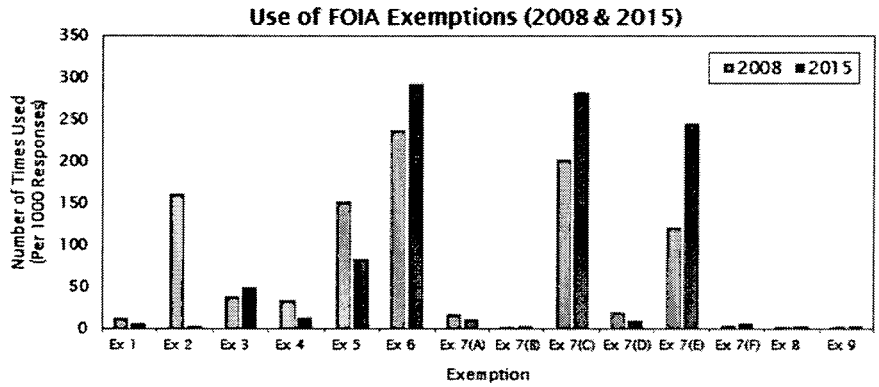
123. *Id.*; *see* U.S. DEP'T OF JUSTICE, GUIDE TO THE FREEDOM OF INFORMATION ACT 417 (2014), https://www.justice.gov/sites/default/files/oip/legacy/2014/07/23/exemption6_0.pdf [https://perma.cc/WD87-SVZV].

124. *See, e.g.*, *Create a Basic Report*, FOIA.GOV, https://www.foia.gov/data.html [https://perma.cc/N7CF-N67U] (last visited Apr. 27, 2017) (select Department of Commerce, Department of Education, Department of Health and Human Services, and the filters of "Exemptions" and "FY 2015," and the generated chart reflects that Exemption 6 is the most often cited reason for failing to release information).

125. *See* Max Galka, *Transparent Censorship: An In-Depth Look at FOIA in 2015*, FOIA MAPPER (Apr. 11, 2016), https://foiamapper.com/annual-foia-reports-2015/ [https://perma.cc/TDU9-YGG8]. Requesting parties may appeal denial decisions in federal court, which reviews denials de novo and generally resolves FOIA disputes at the summary judgment stage. *See, e.g.*, Arieff v. U.S. Dep't of the Navy, 712 F.2d 1462, 1468–69 (D.C. Cir. 1983); Judicial Watch, Inc. v. Dep't of the Navy, 25 F. Supp. 3d 131, 136 (D.D.C. 2014). "If, however, the record leaves substantial doubt as to the sufficiency of the search, summary judgment for the agency is not proper." *See* Truitt v. Dep't of State, 897 F.2d 540, 542 (D.C. Cir. 1990). If the agency is able to provide responsive records, but a portion of the record should be properly withheld, the agency may not deny complete disclosure if the record can be segregated such that the exempt portions are redacted and the nonexempt portions are disclosed unless it is impossible to separate the two. *See* 5 U.S.C. § 552(b).

126. *See* Galka, *supra* note 125.

Figure 1.[127]



Exemption 6 is the leading exemption agencies use to avoid the release of "personnel and medical files and similar files the disclosure of which would constitute a 'clearly unwarranted invasion of personal privacy.'"[128] The Supreme Court determined that the phrase "similar files" within Exemption 6 includes data that, if released, would subject someone to "injury and embarrassment that can result from the unnecessary disclosure of personal information."[129] Only the balancing of the public interest in its right to know against a person's privacy interest in the particular data governs whether the data should be released, and not the type of file itself.[130] Further, exemptions—including those involving privacy—are to be narrowly construed in favor of release, and an agency must distinguish between a substantial and a *de minimis* privacy interest.[131] Examples of information that does not implicate privacy concerns include information not linked to a particular individual and federal employee information that is not personal in nature.[132] Despite these constraints, data privacy serves as an easy excuse for the agency to decline the request.

---

127. *Id.*

128. *See, e.g.*, U.S. Dep't of State v. Wash. Post Co., 456 U.S. 595, 596 n.1, 602–03 (1982) (quoting 5 U.S.C. § 552(b)) (upholding under Exemption 6 the government's denial of requests for documents regarding whether particular Iranian nationals held valid US passports).

129. *Id.* at 599.

130. *Id.*

131. *See* Multi Ag Media LLC v. Dep't of Agric., 515 F.3d 1224, 1230 (D.C. Cir. 2008) (noting "a privacy interest may be substantial—more than *de minimis*—and yet be insufficient to overcome the public interest in disclosure").

132. *See* Arieff v. U.S. Dep't of the Navy, 712 F.2d 1462, 1467 (D.C. Cir. 1983); *see also* Aguirre v. SEC, 551 F. Supp. 2d 33, 54 (D.D.C. 2008) (distinguishing accessible information that "identifies government employees" from properly withheld information related to employment termination or personal travel). The agency bears the burden of demonstrating that the exemption is appropriate. *See* Fed. Open Mkt. Comm. v. Merrill, 443 U.S. 340, 352 (1979); Murphy v. Exec.

Consider, for example, a state law case: the *Southern Illinoisan*'s open records request for information from the Illinois Department of Public Health about the incidence of neuroblastoma in Illinois from 1985 to the date of the request.[133] The agency denied the newspaper's request because it determined the release would violate individual privacy in the database, as researchers would be able to reverse engineer the released data with other publicly available data, identify the individuals by zip code, and determine whether they had cancer.[134] On appeal, the Illinois Supreme Court found that denying release of public data simply because someone could combine outside data with the released data to determine a person's identity was an inappropriate litmus test.[135] Without data from the Illinois Department of Public Health, and given the data restrictions imposed on hospitals and other health organizations by the Health Insurance Portability and Accountability Act (HIPAA), it would have been practically impossible to obtain these data from any other source.[136] Had the court not ordered the release, the only practical solution would have been to drop the investigation altogether.[137]

Rather than unnecessarily refusing to release data based on specific privacy concerns, agencies should be instructed by the US Court of Appeals for the DC Circuit's decision in *Arieff v. Department of the Navy*, finding the Navy's refusal to release general prescription data regarding six hundred patients based on privacy concerns to be inappropriate.[138] Requesting parties asked the Navy for "all records concerning releases of any prescription drugs" between certain time

---

Office for U.S. Attorneys, 789 F.3d 204, 209 (D.C. Cir. 2015) ("We have emphasized that an agency's task is not herculean. The justification for invoking a FOIA exemption is sufficient if it appears logical or plausible." (internal quotations omitted)). A requesting party must not only show "more than a bare suspicion" that the agency acted negligently in denying disclosure but also evidence to produce a reasonable belief of the alleged impropriety's occurrence. *See* Nat'l Archives & Records Admin. v. Favish, 541 U.S. 157, 174 (2004).

133. *See* S. Illinoisan v. Ill. Dep't of Pub. Health, 844 N.E.2d 1, 3 (Ill. 2006). Although a state law case, state open records laws are similar to FOIA and its exemptions. *Compare* 5 U.S.C. § 552(b) (2012), *with* 5 ILL. COMP. STAT. 140/7.5 (2017).

134. *S. Illinoisan*, 844 N.E.2d at 4.

135. *Id.* at 19–20.

136. *See id.* at 18.

137. The court made an interesting observation regarding this situation:

> The entire purpose of the Cancer Registry Act would be effectively repealed by subsection 4(d) if we did not impose the reasonableness requirement, because any fact, no matter how unrelated to identity can tend to lead to identity, and, therefore, any and every fact would be exempt under subsection 4(d).

Id. at 5–6 (internal citation omitted).

138. Arieff v. U.S. Dep't of the Navy, 712 F.2d 1462, 1467, 1471–72 (D.C. Cir. 1983).

periods.[139] The Navy denied the request in part because the release of the data "would constitute a 'clearly unwarranted invasion of [the] personal privacy' of the Beneficiaries in violation of [Exemption 6]."[140] The DC Circuit disagreed, noting any "secondary effect" of the release is irrelevant regarding whether the government FOIA release should occur.[141] Rather, the actual production of the documents must be the cause of the invasion of privacy for the withholding to be proper.[142] In other words, unless a person's identity will clearly be revealed because of the agency's release of the data, rather than mere speculation that the release might identify an individual, the government should release the data. Failing to do so diminishes the sources of data for public researchers and increases the power of data brokers and the privatization of research, leading to data inequality and negative rent seeking.

B. Individual Protections Under the Privacy Act

In contrast to FOIA—a statute of release—the Privacy Act of 1974[143] is a statute of protection from release and is another hurdle for researchers to overcome. The Privacy Act governs "the collection, maintenance, use, and dissemination of personal information by Federal agencies."[144] An individual may access and correct his own information contained within the federal government's records.[145] Any information about the individual that is "linked to that individual by name or identifying particular" is protected from government release.[146] Thus, where a FOIA exemption would permit the government to deny release of personal information, a requester may force its release only when the information pertains to *himself* rather than to third parties.[147]

---

139. *Id.* at 1465.

140. *Id.* (first alteration in original) (quoting 5 U.S.C. § 552(b)(6) (2012)).

141. *Id.* at 1468.

142. *Id.*

143. Privacy Act of 1974, Pub. L. No. 93-579, § 3, 88 Stat. 1896, 1897–910 (1974) (codified as amended at 5 U.S.C. § 552a (2012)).

144. *Id.* § 2(a), 88 Stat. at 1896.

145. 5 U.S.C. § 552a(d)(1) (2012).

146. Pierce v. Dep't of the U.S. Air Force, 512 F.3d 184, 191–92 (5th Cir. 2007) (citing 5 U.S.C. § 552a(a)(4)). A record is

> any item, collection or grouping of information about an individual that is maintained by an agency, including, but not limited to, his education, financial transactions, medical history, and criminal or employment history and that contains his name, or the identifying number, symbol, or other identifying particular assigned to the individual such as a finger or voice print or photograph.

5 U.S.C. § 552a(a)(4).

147. 5 U.S.C. § 552a(d).

As for requests for a third party's information, the Privacy Act prohibits federal agencies from disclosing personally identifiable information without the subject of the request's express consent.[148]

In those instances under FOIA that *require* the government to release information, the Privacy Act likewise *permits* the government to release the information.[149] However, Exemption 6 of FOIA indicates the government's release of personal information is purely discretionary and not mandatory; thus, the Privacy Act would allow the government to withhold information falling within Exemption 6's parameters.[150] Accordingly, where there would be a "clearly unwarranted invasion of privacy" by the release of data under FOIA Exemption 6, the records may not be released without the subject individual's consent under the Privacy Act.[151] In this regard, agencies are left with unfettered discretion to release or not to release data.[152] The general practice is for agencies to decline to release records even where release might be possible.[153] That practice makes researchers' use of free government data more difficult and contributes to data inequality.

Moreover, outside the context of FOIA, the manner in which the government provides access to open data is the subject of administrative policy goals rather than specific laws.[154] Academics, legal scholars, lobbyists, and companies often try to influence the

148. *Id.* § 552a(b).

149. *See, e.g., id.* § 552a(b)(2).

150. *Id.* § 552(b)(6); *see id.* § 552a(b)(2).

151. *Id.* § 552(b)(6); *see id.* § 552a(b)(2).

152. A plaintiff's challenge to the government's release under the Privacy Act must demonstrate that "(1) the information is a record within a system of records; (2) the agency disclosed the information; (3) the disclosure adversely affected the plaintiff; and (4) the disclosure was willful or intentional." Pierce v. Dep't of the U.S. Air Force, 512 F.3d 184, 186 (5th Cir. 2007); *see* Luster v. Vilsack, 667 F.3d 1089, 1097 (10th Cir. 2011). For the government to be held liable, the release of data must have been "so patently egregious and unlawful that anyone undertaking the conduct should have known it unlawful." Maydak v. United States, 630 F.3d 166, 180 (D.C. Cir. 2010); *see* 5 U.S.C. § 552a(g)(4). Third parties (nongovernmental officials) who release private data are not subject to the Privacy Act. Nat'l Labor Relations Bd. v. Vista Del Sol Health Servs., Inc., 40 F. Supp. 3d 1238, 1268 (C.D. Cal. 2014); *see* 5 U.S.C. § 552a(b). Moreover, if the released data are available elsewhere rather than merely contained within the federal government's records, there is no Privacy Act violation if the government releases the same information. *See* York v. McHugh, 850 F. Supp. 2d 305, 310–12 (D.D.C. 2012); *see also* Doe v. U.S. Dep't of Treasury, 706 F. Supp. 2d 1, 6 (D.D.C. 2009) (noting the Privacy Act only applies to direct or indirect releases of data from a governmental system of records).

153. Plaintiffs about whom data is sought may pursue a "reverse FOIA" claim, seeking protection from the release of their data. *See* Doe v. Veneman, 380 F.3d 807, 810 (5th Cir. 2004).

154. *See generally id.*

executive branch's policies.[155] Access to the executive branch potentially enables individuals and interest groups to influence existing and future policy.[156] Whether and how data giants have access to the current presidential administration could dictate the likelihood the public and researchers will have access to free, unbiased, and fact-checked data, rather than data for purchase or subject to trade secret protections. If future administrations allow commercial data brokers to dictate government policy, it will further data inequality and weaken research credibility as researchers will turn to data brokers for their information. Only those researchers with adequate funding or those who collaborate with data brokers will contribute to the future of research.

C. Government's De-Identification of Data and Concomitant Privacy Concerns Unreasonably Dominate Its Release Decisions

The Authors contend that generalized release denials that lack consideration of a person's voluntary revelation of the same data on the Internet will promote further data inequality. Nonetheless, there are countervailing policy considerations associated with exposing one's personal data to the government, including (1) concern over third parties accessing private data through a FOIA request and (2) concern over the efficacy of anonymization techniques used in the government's general release of statistics leading to revelation of private data. Many government services require personal data to utilize that service.[157] Initially, there may be a chilling effect and disincentive for individuals to provide the government with personal information, knowing the government may store it and that it may be subject to release.[158] The possibility of re-identification is significant particularly because data are no longer within the individual's control.[159] The data can be subject

155. One entity, Google, had the most significant contact with the Obama administration in small groups or individual meetings with key White House officials. Between January 2009 and October 31, 2015, Google met at the White House approximately 427 times. *See* Kampis, *supra* note 29. This exceeds the number of meetings that all top fifty oil and gas companies had with the White House during the same time frame. *See id.* Because the White House is not subject to FOIA, however, whether the visitor logs actually capture all meetings is unclear. *Id.* Prior administrations did not make visitor logs publicly available, and there is no obligation that future administrations do so. *Id.* Indeed, the Trump administration has indicated it will not release visitor logs. *See* Julie Hirschfeld Davis, *White House to Keep Its Visitor Logs Secret*, N.Y. TIMES (Apr. 14, 2017), https://www.nytimes.com/2017/04/14/us/politics/visitor-log-white-house-trump.html?_r=1 [https://perma.cc/CE4W-NH8U].

156. *See* Altman et al., *supra* note 74, at 1999.

157. *See* Borgesius et al., *supra* note 20, at 2088.

158. *Id.*

159. *Id.* at 2091.

to misuse or abuse.[160] Because of these privacy concerns, individuals may be disinclined to inquire about services for issues that are relevant to the public health sector like pregnancy, disease, drugs, financial issues, or suicidal thoughts.[161] Although the chilling effect certainly can impact the individual, society as a whole is likewise impacted.[162] Privacy violations because of the government's data release have been few and far between.[163] Government agencies have done an admirable job of balancing the need for privacy while also providing the public with statistical data.[164] These agencies have been at the forefront of developing tools and techniques to make this possible.[165]

For example, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)[166] governs the government's release of statistical data.[167] CIPSEA's terms dictate how the federal government can prevent identification of an individual through the public release of statistics when aggregated from a variety of governmental sources involving that individual.[168] In response to privacy concerns, agencies utilize de-identification[169] tools known as "statistical disclosure limitation techniques"[170] to prevent individuals' identification.[171] The purpose of statistical disclosure limitation

160. *See id.* at 2088–92.

161. *Id.* at 2088.

162. *Id.* at 2088–89.

163. *See* Altman et al., *supra* note 74, at 2001; *see also id.* at 1985 ("One appeals court, for instance, held that an agency's negligent actions did not violate the law even though the trial court had found that the privacy violations had been 'substantial.'").

164. *Id.* at 1993–94.

165. *Id.*

166. Confidential Information Protection and Statistical Efficiency Act of 2002, Pub. L. No. 107-347, §§ 501–26, 116 Stat. 2899, 2962–70 (2002) (codified as amended at 44 U.S.C. §§ 3501–21 (2012)).

167. *See* Altman et al., *supra* note 74, at 1992.

168. *Id.* at 1993–94.

169. "De-identification" is a process or set of processes that utilizes a variety of tools to mask and prevent the ability of a third party from identifying any one particular individual from an aggregated data set. *See* SIMSON L. GARFINKEL, NAT'L INST. OF STANDARDS & TECH., NISTIR 8053, DE-IDENTIFICATION OF PERSONAL INFORMATION 1 (2015), http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf [https://perma.cc/V3QV-K2L2]. Anonymization is also a method of preventing the identification of an individual in a data set, and the identity of that individual remains unknown to the collector as well. For de-identification, by contrast, the identity of the individual may be known to the collector. *Id.* at 2–3.

170. *See* Altman et al., *supra* note 74, at 1972–73; *see also* Ira S. Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. 703, 712–13, 717 (2016).

171. Altman et al., *supra* note 74, at 2004. Many agencies have disclosure review boards or panels to ensure release does not breach privacy rights. *Id.* at 1994. A few months prior to the proposed release, the agencies compare their disclosure limitation techniques with the availability of similar data potentially linked to the proposed release data. *Id.* Other statistical disclosure laws may apply depending on the agency. *Id.*

techniques is to prevent the disclosure of an individual's identity or personal attributes when data are released to the public in an aggregate form or released to researchers in the form of microdata.[172] These techniques include redacting personal identifiers, coarsening attributes such as modifying a person's location, recoding the values associated with a person into rounded values or intervals, swapping values in similar records, truncating extreme values, and adding random noise to the data.[173] These tools add background noise to the statistical data, making it more difficult to accurately identify a particular person in any aggregated materials.[174] These tools consider the impact the modifications have on the utility of the data as well as the extent to which they prevent unwarranted disclosure.[175] However, these tools do not take into consideration whether the subject has otherwise provided the information to a data broker. The Authors contend this should be added as a consideration for analysis. If individuals freely provide information—which the government likewise possesses—the type of information, and to whom the information was provided, should be part of the government's assessment as to the sensitivity of the information. In those instances where the information is freely provided to a variety of online sources and the information does not relate to issues of identity theft or other sensitive information, the release might be appropriate with minimal statistical limitation techniques.

Consider the case of National Center for Science and Engineering Statistics (NCSES). As part of its mission, NCSES conducts a survey of doctoral recipients and makes this data available to the public through its tabulation engine.[176] The website notes that "[t]he tabulation engine includes a disclosure control mechanism that protects the identity of respondents when using the gender, citizenship, and race/ethnicity variables."[177] A request for data regarding the race and ethnicity of the computer science PhD graduates at the University of North Texas provides the information demonstrated in Figure 2.[178]

---

172. *See id.* at 1972–73.

173. *Id.* at 1995.

174. *Id.* at 1972–73.

175. *Id.* at 1973. The impact of de-identification is relevant because it decreases data's utility. *See id.* at 1973–74; *see also* Rubinstein & Hartzog, *supra* note 170, at 709–10 (noting the failure of anonymization technology to protect privacy leading to polarization between policy makers).

176. *Welcome*, NAT'L CTR. FOR SCI. & ENGINEERING STAT., https://ncses.norc.org/NSFTabEngine/#WELCOME [https://perma.cc/T7EQ-ZPC6] (last visited Feb. 8, 2018).

177. *Id.*

178. *Tabulation*, NAT'L CTR. FOR SCI. & ENGINEERING STAT., https://ncses.norc.org/NSFTabEngine/#TABULATION (last visited Feb. 15, 2018) (Outer Row:

Figure 2.[179]

Year	Academic Discipline, Detailed (standardized)	Hispanic of Latino	American Indian or Alaska Native, non-Hispanic	Asian, non-Hispanic	Black, non-Hispanic	White, non-Hispanic	Two or more races, non-Hispanic	Ethnicity not reported	Total
2015	Computer Science	***	***	***	***	11	***	3	21

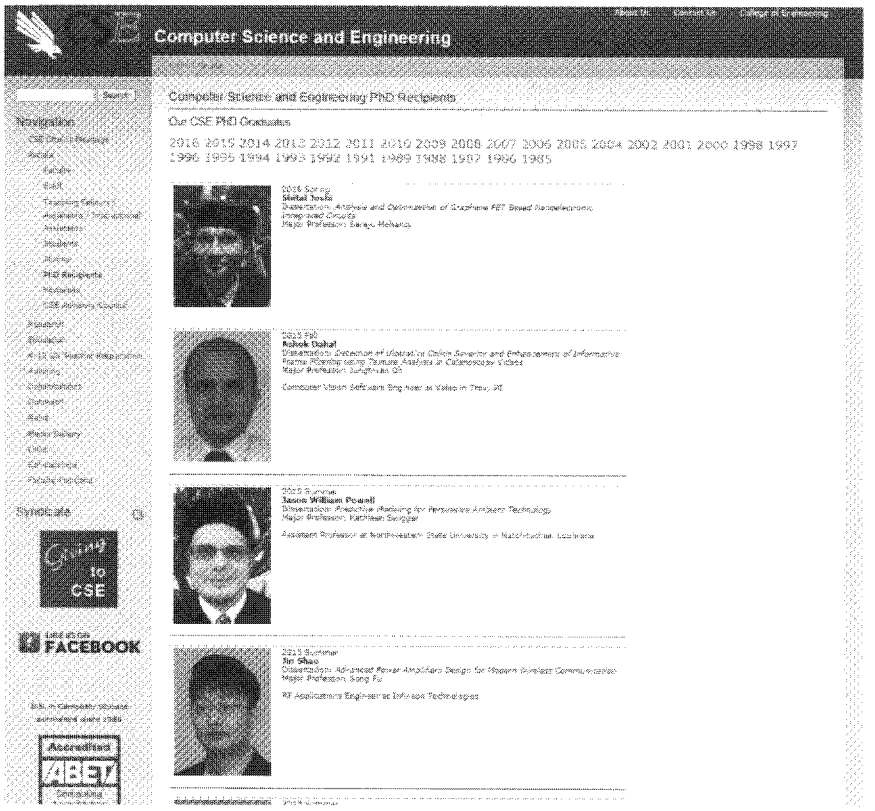
For a researcher conducting research on minorities receiving a PhD in computer science, the figure above is practically worthless. At the same time, the University of North Texas lists all of its computer science doctoral recipients on its public web site.[180] A screenshot of the website is shown in Figure 3.

Year; Inner Row: Discipline, Detailed; Column: Race/Ethnicity; Filter: Institution (STD), input "University of North TX," and select the Denton campus).

179. The "***" represents information that was suppressed to prevent the disclosure of the identity of the individuals who received a doctorate. *Id.* In addition, "other cells will also be suppressed (secondary suppression) to protect those initial cells from being mathematically calculated." *Frequently Asked Questions*, NAT'L CTR. FOR SCI. & ENGINEERING STAT., https://ncses.norc.org/NSFTabEngine/#HELP_FAQ [https://perma.cc/L4F9-LCXN] (last visited Feb. 9, 2018).

180. *See PhD Graduates*, COMPUTER SCI. & ENGINEERING, U. NORTH TEX., http://computerscience.engineering.unt.edu/phd-graduates [https://perma.cc/KS9H-ESVZ] (last visited Feb. 9, 2018).

Figure 3.[181]



Arguably, information about the ethnicity of the doctoral recipients in computer science from the University of North Texas can be speculated from the names and photos of the PhD graduates listed on the university's website. Even when a university does not publicize doctoral recipients to this extent, it is likely that this information is readily accessible from other sources within the university (e.g., graduation lists) or even from the doctoral recipient's website; yet NCSES suppresses this information when it releases summary data. In this regard, the Authors contend privacy concerns unreasonably dominate the government's release decisions.

Further concern over the public's ability to combine discrete data sets in a manner that then identifies an individual has led the government to analyze the released data using a "mosaic effect."[182] In

181. *Id.* This screenshot was taken on September 15, 2017, but the website has since updated its list of recipients.

182. ComputerWorld defines the mosaic effect as "[d]ata elements that in isolation look relatively innocuous [but] amount to a privacy breach when combined." *See* Jaikumar Vijayan,

determining whether to release statistical or other mass data points, government agencies assess the impact the release will have on an individual's personal data and utilize a conservative approach to prevent disclosure of sensitive data.[183] The assessment, discussed more fully in the next Part, is known as a "privacy first" assessment that the Authors contend will lead to the further privatization of research if not balanced with an assessment of voluntary revelation by the individual himself. It seems completely counterproductive to move the entire responsibility of protecting privacy to the government while allowing data brokers to operate entirely without restrictions. If the government is prevented from releasing the type of data that the data brokers are free to sell, then the inevitable result is that the data brokers will corner the market on data, resulting in another form of negative rent seeking.

V. THE PUBLIC'S ABILITY TO ACCESS DATA FROM DATA BROKERS AND THE DATA BROKERS' PRIVACY OBLIGATIONS

Despite significant pressures on the government to ensure data privacy, a consistent regulation for maintaining the privacy of data gathered, distributed, or maintained by private entities has not emerged.[184] This unusual discrepancy between strong privacy obligations for the government and nearly nonexistent privacy obligations for the data brokers furthers a negative rent-seeking situation. In 1989, the Supreme Court stated: "[T]he common law and the literal understandings of privacy encompass the individual's control

---

Sidebar: The Mosaic Effect, COMPUTERWORLD (Mar. 15, 2004, 12:00 AM), http://www.computerworld.com/article/2563635/security0/sidebar--the-mosaic-effect.html [https://perma.cc/YW6T-GGUV]. Data experts agree that there is no foolproof way to ensure that disclosure limitation techniques will eliminate the ability of a third party to cull together data and identify a particular individual within a discrete data set. *See* Altman et al., *supra* note 74, at 1973–74. Both Netflix and America Online released anonymized data that others were able to compare with publicly available data to identify those members contained within their studies. RUMSEY, *supra* note 88, at 7

183. *See* RUMSEY, *supra* note 88, at 1–2; Altman et al., *supra* note 74, at 2001. Unlike public data sets, restricted data sets require researchers to apply for access to the data, and the governmental release depends on a formal screening process. Altman et al., *supra* note 74, at 1996. The use is limited to the purposes specified through data-use agreements. *Id.* Regarding data related to individuals (financial, demographic, purchasing behavior, etc.), if the government agency refuses to release data, the researcher has the option of purchasing the data from data aggregators; likewise, data related to organizations have long been available from other sources (CRSP, COMPUSTAT, and others). *Id.* at 1995–96. *See generally* Sema Dube et al., *Is Hostility in the Merger and Acquisition Market Wasteful? Empirical Evidence of the Economic Costs of Hostility*, 7 J. BUS. & SEC. L. 9, 19 (2007) (utilizing the CRSP and COMPUSTAT databases to obtain market-related data).

184. *See* Paul Ohm, *Sensitive Information*, 88 S. CAL. L. REV. 1125, 1138–39 (2015).

of information concerning his or her person."[185] However, the right to privacy outlined in the US Constitution merely prevents the government from "intrusive government activities."[186] It does not protect individuals from private-sector intrusion.[187] In the United States, courts treat personal data as a product rather than a right, as opposed to the European Union, which considers these rights fundamental.[188] As such, a more balanced approach in favor of relaxing the government's privacy obligations along with an increase in access to data broker information is necessary.

A. Access to Public Data

In light of the government's declination to release data, or to release only thoroughly scrubbed data, researchers have better access to purchased data.[189] Commercial entities have found data analytics to be a big business.[190] Indeed, anyone can buy just about any data from a data broker.[191] Researchers have more access to privately gathered data, and commercial vendors of this type of research data are growing.[192] Commercial vendors include Datasift, Acxiom, Treato, and

185. *See* U.S. Dep't of Justice v. Reporters Comm. for Freedom of Press, 489 U.S. 749, 763 (1989).

186. *See* Kuempel, *supra* note 32, at 214.

187. *Id.*

188. *Id.* at 215.

189. Marketing of data gathered by data brokers accounts for the largest amount of their revenue generation, followed by risk-mitigation and people-search products. *See* 2014 DATA BROKER REPORT, *supra* note 12, at 23. Data brokers often require their clients to certify that they will not violate a federal law like the Fair Credit Reporting Act. *Id.* at 16–17. However, the data brokers do not monitor or review whether violations occur. *See id.* The only limitation on what can be purchased is the efficacy of the particular data broker. Data brokers dictate the nature of the relationship with the consumer through standard contractual agreements. *Id.* at 16. Data brokers and their sources generally enter into one of three types of contractual relationships with purchasers: (1) outright ownership of the gathered data, (2) license to use the data for a certain time period, or (3) the right to resell the data. *Id.*

190. *See* Breggin & Amsalem, *supra* note 18, at 10986; Kuempel, *supra* note 32, at 209–10.

191. *See* Caitlyn Renee Miller, *I Bought a Report on Everything That's Known About Me Online*, ATLANTIC (June 6, 2017), https://www.theatlantic.com/technology/archive/2017/06/online-data-brokers/529281/ [https://perma.cc/78AV-5LC8].

192. *See* Mattioli, *supra* note 24, at 558. When asked about its clients and data sources during a Senate investigation, data broker Acxiom refused to reveal this data but generically noted it works for "47 Fortune 100 clients," "5 of the 13 largest U.S. federal government agencies," and "[b]oth major national political parties." *See* Gregory Maus, *How Corporate Data Brokers Sell Your Life, and Why You Should Be Concerned*, STACK (Aug. 24, 2015, 2:27 PM), https://thestack.com/security/2015/08/24/how-corporate-data-brokers-sell-your-life-and-why-you-should-be-concerned/ [https://perma.cc/3E5T-AB7V].

TrueLens.[193] Data brokers collect data from a variety of sources, including public records, loyalty cards, websites, social media, bankruptcy data, voting history, consumer purchase data, web browsing activities, and warranty registrations.[194]

To what extent privacy interests are protected from a data broker's release of information is governed by one main regulatory body—the FTC—which has authority over consumer data brokers under section 5 of the Federal Trade Commission Act.[195] In 2014, the FTC issued a report regarding data brokers, entitled *Data Brokers: A Call for Transparency and Accountability* (the "Data Broker Report").[196] In the Data Broker Report, the FTC addressed (1) marketing products, (2) risk-mitigation products, and (3) people-search products offered by data brokers.[197] The Data Broker Report noted that consumers benefit from easier access to goods and services and to lower-cost or free web services because these services derive financial benefits from consumers through the sale of specifically marketed advertisements

193. *See* Meta S. Brown, *16 Major Data Vendors*, DUMMIES, http://www.dummies.com/programming/big-data/16-major-data-vendors/ [https://perma.cc/2JNR-C97U] (last visited Feb. 9, 2018); *Treato*, CRUNCHBASE, https://www.crunchbase.com/organization/treato [https://perma.cc/A95V-FEPX] (last visited Feb. 9, 2018); *TrueLens*, CRUNCHBASE, https://www.crunchbase.com/organization/truelens (last visited Feb. 9, 2018).

194. 2014 DATA BROKER REPORT, *supra* note 12, at iv.

195. *See* Federal Trade Commission Act, ch. 311, § 5, 38 Stat. 717, 719 (1914) (codified as amended at 15 U.S.C. § 45(a)(2) (2012)). In 2016, the Federal Communications Commission (FCC) attempted to join the field of privacy regulation for telephone and cable companies by enacting the Protecting the Privacy of Broadband and Other Telecommunications Services Order (the "Privacy Order"). *See* Protecting the Privacy of Broadband Customers, 31 FCC Rcd. 13911 (2016), *superseded by* Restoring Internet Freedom, 83 Fed. Reg. 7852 (Feb. 22, 2018) (to be codified at 47 C.F.R. pts. 1, 8, 20). The Privacy Order was designed to limit the quantity of data a telephone or broadband provider collects about its consumers, including their "geo-location; health, financial, and children's information; Social Security numbers; content; and web browsing and application usage histories[.]" *Id.* at 13977. Consumers would have had to consent to the Internet service providers' use and sharing of such data for anything other than the purposes for which the broadband provider services the consumer—for example, billing. *See id.* at 13959–60. However, under the new administration, these regulations have been revoked and are unlikely to be implemented any time soon. *See* Alina Selyukh, *As Congress Repeals Internet Privacy Rules, Putting Your Options in Perspective*, NPR (Mar. 28, 2017, 6:58 PM), https://www.npr.org/sections/alltechconsidered/2017/03/28/521813464/as-congress-repeals-internet-privacy-rules-putting-your-options-in-perspective [https://perma.cc/D9XL-GAMK] (detailing ways consumers can protect their own privacy online and on smartphones). For an in-depth discussion about the Restoring Internet Freedom Rule, see Rob Frieden, *Freedom to Discriminate: Assessing the Lawfulness and Utility of Biased Broadband Networks*, 20 VAND. J. ENT. & TECH. L. 655 (2018).

196. *See* Kuempel, *supra* note 32, at 234; *see also* 2014 DATA BROKER REPORT, *supra* note 12, at i (defining the term "data broker" as a company that "collect[s] consumers' personal information and resell[s] or share[s] that information with others").

197. 2014 DATA BROKER REPORT, *supra* note 12, at 23.

based on the data brokers' data.[198] Nonetheless, the Data Broker Report noted areas for improvement: (1) the need for transparency in data brokers' policies, (2) the "aggregation effect" leading to potentially discriminatory use of data, and (3) the potential security risks with stored data.[199] The FTC studied nine data brokers, representing over one thousand companies.[200] Acxiom, one of the nine data brokers the FTC studied, has three thousand discrete data segments for nearly every US consumer.[201] Data brokers gather and aggregate data into discrete categories, identifying consumers as, *inter alia,* the "expectant parent," "bible lifestyle," and "financially challenged."[202]

Data brokers often gather data from the government. For example, the Census Bureau issues demographic studies that identify "ethnicity, age, education level, household makeup, income, occupations, and commute times," along with "geographic information including roads, addresses, congressional districts, and boundaries for cities, counties, subdivisions, and school and voting districts."[203] The Social Security Administration's Death Master File lists "consumers' names, [social security numbers], and dates of death," and the US Postal Service discloses address standardization and change of address data.[204] Additional governmental data is provided by state agencies, such as licensing records, real property records, taxes, voter registration, court records, and motor vehicle records.[205] The data brokers also filter social media and other Internet blogs and posts, garnering data when the user does not set privacy restrictions.[206]

198. *Id.* at 47.

199. *See* Kuempel, *supra* note 32, at 218–22; *see also* 2014 DATA BROKER REPORT, *supra* note 12, at 51–52. The FTC recommends four basic areas in need of legislative action, including (1) requiring data brokers to provide consumers with access to their data, including sensitive data to a reasonable level of detail; (2) allowing consumers the option of opting out of having the data shared for marketing purposes; (3) informing consumers of the source of their data so that they can correct any inaccurate information; and (4) obtaining a consumer's prior affirmative consent where sensitive data is being collected. 2014 DATA BROKER REPORT, *supra* note 12, at viii.

200. *See* 2014 DATA BROKER REPORT, *supra* note 12, at ii; Kuempel, *supra* note 32, at 211. The nine data brokers who received FTC requests for data were: Acxiom, Corelogic, Dataogix, eBureau, ID Analystics, Intelius, PeekYou, Rapleaf, and Recorded Future.

201. *See* 2014 DATA BROKER REPORT, *supra* note 12, at 8 ("Acxiom provides consumer data and analytics for marketing campaigns and fraud detection. Its databases contain information about 700 million consumers worldwide with over 3000 data segments for nearly every U.S. consumer."); *see also* Maus, *supra* note 192.

202. Maus, *supra* note 192.

203. *See* 2014 DATA BROKER REPORT, *supra* note 12, at 11.

204. *Id.* Other entities, including the Federal Bureau of Investigation, US Secret Service, and the European Union, provide information "related to terrorist watch lists or most wanted lists." *Id.*

205. *Id.*

206. *Id.* at 13.

Finally, data brokers purchase data from commercial data sources, such as retailers learning a consumer's purchase histories, purchase prices, dates of purchase, and form of payment used, along with registration sites such as news and travel sites.[207] Privacy concerns surrounding data brokers gathering and selling consumer data are apparent, but regulation is limited.

With respect to these private entities, the origins of privacy law are based in tort or are contained within discrete sets of specific legislation, such as laws protecting driver's license data and credit reports.[208] The release of nonsensitive data does not necessarily result in harm, but others could use it to de-identify data that used anonymization tools to protect consumer privacy.[209] Unfortunately, once the data broker or private entity sells the data to a third party, the FTC's jurisdiction likely ceases.[210] The FTC can regulate when a business sets a privacy policy or markets the privacy of its product and the practice is found to be deceptive.[211] However, the FTC cannot require that a company set a privacy policy.[212] Those that do not have privacy policies and do not promise privacy to their customers are exempt from liability except for common law privacy tort claims.[213]

---

207. *See id.*

208. *See* Ohm, *supra* note 14, at 1732–35. In addition to the initiatives surrounding government records and their transparency, the Obama administration attempted, unsuccessfully, to rectify some of the concerns over individual privacy rights and commercial data brokers through the Consumer Privacy Bill of Rights in 2012. *See* Natasha Singer, *Why a Push for Online Privacy Is Bogged Down in Washington*, N.Y. TIMES (Feb. 28, 2016), https://www.nytimes.com/2016/02/29/technology/obamas-effort-on-consumer-privacy-falls-short-critics-say.html?_r=0 [https://perma.cc/PW95-ZXDJ]. Likewise, the Data Broker Accountability and Transparency Act of 2015 and the Data Security and Breach Notification Act, which would increase consumer rights, have repeatedly failed in committee. *See* Maus, *supra* note 192.

209. *See* Amelia R. Montgomery, Note, *Just What the Doctor Ordered: Protecting Privacy Without Impeding Development of Digital Pills*, 19 VAND. J. ENT. & TECH. L. 147, 157–58 (2016).

210. *See* Kwame N. Akosah, Note, *Cracking the One-Way Mirror: How Computational Politics Harms Voter Privacy, and Proposed Regulatory Solutions*, 25 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 1007, 1045–46 (2015). Other privacy laws may assist in regulating big data's use of certain data. *See id.*; Eric Everson, *Privacy by Design: Taking Ctrl of Big Data*, 65 CLEV. ST. L. REV. 27, 37 (2016) (noting the wide array of federal and state laws targeting privacy issues).

211. *See* Lipman, *supra* note 113, at 790–92.

212. *Id.*

213. Michael S. Yang, *E-Commerce: Reshaping the Landscape of Consumer Privacy*, 33 MD. B.J. 12, 14 (2000); *see* Lipman, *supra* note 113, at 789. However, the FTC has had some success in this area with one administrative action against Google for its misrepresentation of what it collected when consumers utilized Apple's Safari Internet browser. *See* Baranetsky, *supra* note 15, at 331–32. In another FTC administrative compliance case, the FTC found Facebook deceived its users by allowing their data concerning those items they marked with a "like" to be public even though Facebook led its users to believe such data were private. *Id.* Whether the Defend Trade Secrets Act of 2016, an amendment to the Economic Espionage Act, will provide any additional privacy rights to individuals is yet to be seen. *See, e.g.*, Defend Trade Secrets Act of 2016, Pub. L.

Interestingly, data brokers are somewhat self-regulated by industry trade associations that have identified best practices for handling consumer data.[214] Although self-regulation provides some guidance for protecting consumer data, it is purely voluntary.[215] A joint task force made up of various governmental and nongovernmental stakeholders—including web application providers—has suggested certain best practices.[216] One suggestion is that providers issue a "data disclosure chart" that would require applications to display the types of data the application collects from the user.[217] Because these best practices are purely voluntary, not all entities implement them.[218] Thus, it is unlikely self-regulation will lead to any meaningful privacy or release practices in this area without some governmental or industry incentives.

B. Conflicting Theories: Governmental Release Versus Private Entity Release of Aggregated Data

In situations where the government agency refuses to release the data for privacy reasons and denies the researcher access to public data, the researcher has few alternatives: he may use surrogate measures, alter the research question, or drop the inquiry altogether.[219] Research outcomes are adversely affected by all three options.[220]

Historically, researchers attempted to strike a balance between the risk of disclosing personal data and the usefulness of the released data when requesting government data that used statistical disclosure techniques.[221] About ten years ago, researchers from Microsoft

---

No. 114-153, § 2(a), (b), 130 Stat. 376, 376, 380 (2016) (codified as amended at 18 U.S.C. §§ 1836(b), 1839(3) (2012)) (providing individuals with a federal private cause of action where the plaintiff has taken all reasonable steps necessary to keep their data secret and the data derives an independent economic value if made generally known to the public). This statute could form the basis of a private action against data brokers' repurposing of consumer data without their knowledge.

214. *See* Kuempel, *supra* note 32, at 216; *see also* Singer, *supra* note 208 (discussing how the industry's self-regulation was designed to work in conjunction with the Department of Commerce). However, lack of consensus has inhibited solid industry self-regulation. *See* Singer, *supra* note 208.

215. *See* Kuempel, *supra* note 32, at 216–17.

216. *See* Singer, *supra* note 208.

217. *Id.*

218. *See, e.g.*, *id.* (discussing use and nonuse of voluntary disclosure notices). Other forms of best practices in data gathering of face-recognition and voice-recognition technologies have failed, stalling any cooperative self-regulatory efforts. *Id.*

219. *See* Rubinstein & Hartzog, *supra* note 170, at 719–20.

220. *See id.* at 724. For more information on what agencies must disclose to the public, see Freedom of Information Act, Pub. L. No. 90-23, 81 Stat. 54 (1967) (codified as amended at 5 U.S.C. § 552 (2012)).

221. *See* Altman et al., *supra* note 74, at 1977.

developed what is known as the "privacy first" or "differential privacy" model, meaning a person's privacy interest is more significant in determining whether to release data than any other consideration.[222] Differential privacy—coined by Cynthia Dwork, a highly respected computer scientist and researcher with Microsoft—is a procedure for assessing the risk resulting from a data release, whether through government release or otherwise.[223] These researchers claim data release should be analogous to encryption considerations.[224] Differential privacy can be summarized as follows: how to release data from a data set consisting of n records so that a malicious user who has access to the true values of $(n - 1)$ of those records would not be able to infer data about the remaining nth record.[225] The researchers contend that the data available from data brokers makes such an inference a likely scenario because the malicious user can purchase data about the $(n - 1)$ records.[226]

One prime example to which the privacy first theorists point is the inadvertent sharing of anonymized data from subscribers of Netflix.[227] As part of a marketing contest, Netflix allowed anyone to register for a chance to win $1 million for creating a movie rating system that was better than its existing system and would provide contestants with a "training data set consist[ing] of more than 100 million ratings from over 480 thousand randomly-chosen, anonymous customers on nearly 18 thousand movie titles."[228] Thereafter, a follow-up contest provided access to additional customer data including zip codes, ages, gender, genre ratings, and previously chosen movies.[229] A group of researchers accessed the data by registering for the contest and, instead of modifying the formula, reverse engineered the data identifying the Netflix customers by comparing the data to publicly available records.[230] The researchers demonstrated that if a person knows a bit

222. Chin & Klinefelter, *supra* note 98, at 1427; Christopher Soghoian, *An End to Privacy Theater: Exposing and Discouraging Corporate Disclosure of User Data to the Government*, 12 MINN. J.L. SCI. & TECH. 191, 191–92 (2011); *see, e.g.*, *Data Policy*, FACEBOOK, https://www.facebook.com/policy [https://perma.cc/59N2-ERNQ] (last updated Sept. 29, 2016).

223. *See* Chin & Klinefelter, *supra* note 98, at 1422 n.17, 1429–30; Mona Lalwani, *Apple's Use of 'Differential Privacy' Is Necessary but Not New*, ENGADGET (June 14, 2016), https://www.engadget.com/2016/06/14/apple-differential-privacy/ [https://perma.cc/CM7M-NQN7].

224. Chin & Klinefelter, *supra* note 98, at 1426–27.

225. *See id.* at 1430.

226. *Id.* at 1422.

227. *See, e.g.*, *id.* at 1424–25; Wu, *supra* note 114, at 1118–20.

228. Wu, *supra* note 114, at 1118–19.

229. *Id.* at 1119.

230. *Id.* at 1119–20. The researchers demonstrated the ease with which a person could be identified despite Netflix's de-identification of customer data by assigning random identifiers and

about someone and their viewing habits—information gleaned from an office colleague's discussions at work, for example—that person could take that data in combination with the data provided by Netflix and determine that particular person's viewing habits.[231] A customer sued Netflix for this breach of her personal privacy and reached a settlement with the company.[232] At its core, Netflix was attempting to release useful data that could be used to further its goal of improving services, yet doing so exposed users' privacy interests even though it anonymized the data sets.[233] Clever individuals were able to use the aggregate data in combination with other sources to re-identify some individuals despite the anonymization technology.[234]

Differential privacy theorists recommend that the public release of data be modified to account for this possibility regardless of what the data reveal (sensitive or otherwise).[235] Certainly, privacy is an important consideration; however, a default rule of privacy first may not be the best method to protect data. Researchers have shown that in some cases differential privacy may result in meaningless data.[236] Also, one of the primary concepts of differential privacy is that all aspects of the protection mechanism must be transparent.[237] Practical implementations of differential privacy have been anything but

---

deliberately "perturbing" the data by "deleting ratings, inserting alternative ratings and dates, and modifying rating dates." *Id.*

231. *Id.* at 1120.

232. *Id.*; Ryan Singel, *Netflix Spilled Your* Brokeback Mountain *Secret, Lawsuit Claims*, WIRED (Dec. 17, 2009, 4:29 PM), https://www.wired.com/2009/12/netflix-privacy-lawsuit [https://perma.cc/YW5F-LNMZ].

233. Wu, *supra* note 114, at 1121.

234. *Id.* Although we do not know the exact reason that Netflix settled the case, one can speculate that Netflix may have been concerned that releasing the data violated the Video Privacy Protection Act. *See, e.g.*, Ann M. Schultz, *Protecting Consumer Viewing Habits: Reflections on the Video Privacy Protection Act*, WAYNE ST. U. BLOGS: INFO. POL'Y FOR EVERYDAY DECISIONS (Nov. 30, 2013), https://blogs.wayne.edu/informationpolicy/2013/11/30/protecting-consumer-viewing-habits-reflections-on-the-video-privacy-protection-act/ [https://perma.cc/JCB5-42H9]; *see also* Singel, *supra* note 232. Additionally, as is often the case, settling a matter may help reduce bad publicity, particularly for Netflix's relatively new streaming services. Unlike the *Southern Illinoisian* case where the Health Department was only contending that re-identification *might* occur, researchers in the Netflix case claimed to have *actually* re-identified some individuals using the Netflix data. Singel, *supra* note 232.

235. *See, e.g.*, Wu, *supra* note 114, at 1121.

236. *See* Jane Bambauer, Krishnamurty Muralidhar & Rathindra Sarathy, *Fool's Gold: An Illustrated Critique of Differential Privacy*, 16 VAND. J. ENT. & TECH. L. 701, 704–07 (2014).

237. *See, e.g.*, Phillip Rogaway, The Moral Character of Cryptographic Work 20–21 (Dec. 2015) (unpublished manuscript), http://web.cs.ucdavis.edu/~rogaway/papers/moral-fn.pdf [https://perma.cc/Z7GN-TT56]. *See generally* Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, 54 COMM. ACM 86 https://www.microsoft.com/en-us/research/wp-content/uploads/2011/01/dwork_cacm.pdf [https://perma.cc/JGR4-SM8J].

transparent.[238] For example, Facebook apparently utilizes a form of "differential privacy" in its advertisement targeting databases, which allow an advertiser to target specific users.[239] However, the details of the methods used are entirely unknown to the general public.[240] Moreover, Apple recently announced it would implement a "differential privacy"–style process without explanation of its methods, and recently such methods have proven questionable.[241] Commentators criticized the failure to divulge how such differential privacy techniques are utilized, noting "[i]n the end, one must compare the reduction in harm actually afforded by using differential privacy with the increase in harm afforded by corporations having another means of whitewash, and policy-makers believing, quite wrongly, that there is some sort of cryptomagic to protect people from data misuse."[242]

This lack of transparency from Facebook and Apple is not surprising. Recently, advertisers such as AT&T and Johnson & Johnson pulled their advertisements from YouTube and Google because they found that their advertisements were appearing on websites that promote hate.[243] One would think that it would be easy for Google, a technology giant, to fix the problem instantaneously. However, the algorithms used to place the advertisements are so complex that Google has not been able to assure the advertisers that their advertisements will not appear on inappropriate websites.[244] Astonishingly, Google

---

238. *See* Andy Greenberg, *Apple's 'Differential Privacy' Is About Collecting Your Data—But Not* Your *Data*, WIRED (June 13, 2016, 7:02 PM), https://www.wired.com/2016/06/apples-differential-privacy-collecting-data/ [https://perma.cc/26G2-9QWS]; *see also* Matthew Green, *What Is Differential Privacy?*, BLOG: A FEW THOUGHTS ON CRYPTOGRAPHIC ENGINEERING (June 15, 2016), https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/ [https://perma.cc/S5JP-8GGP].

239. *See generally* Chin & Klinefelter, *supra* note 98, at 1432. *But see* Bambauer, Muralidhar & Sarathy, *supra* note 236, at 738 ("In one case, legal scholars jumped to the conclusion that Facebook employs differential privacy when it is very likely using a different noise-adding technique.").

240. Chin & Klinefelter, *supra* note 98, at 1433.

241. *See* Greenberg, *supra* note 238; *see also* Green, *supra* note 238; Andy Greenberg, *How One of Apple's Key Privacy Safeguards Falls Short*, WIRED (Sept. 15, 2017, 9:28 AM), https://www.wired.com/story/apple-differential-privacy-shortcomings/ [https://perma.cc/MQJ4-P2ZD] (discussing researchers' findings that Apple uploads more specific data than a differential privacy advocate would consider private and protects its methods, which could further eliminate a person's privacy protections with no ramifications).

242. *See* Rogaway, *supra* note 237, at 21.

243. *See* Sapna Maheshwari & Daisuke Wakabayashi, *AT&T and Johnson & Johnson Pull Ads from YouTube*, N.Y. TIMES (Mar. 22, 2017), https://www.nytimes.com/2017/03/22/business/atampt-and-johnson-amp-johnson-pull-ads-from-youtube-amid-hate-speech-concerns.html?mtrref=www.google.com&gwh=5D7AF9F61682EC002C5B62FFA16C3599&gwt=pay [https://perma.cc/4XXQ-Q7FA].

244. *Id.*

itself seems to have been unaware of this problem.[245] The use of complex algorithms whose inner workings that cannot be comprehended easily is a basic way for technology giants to build a smoke screen to protect their operations from the public.[246] The use of opaque differential privacy is consistent with this approach, lulling the public into feeling secure and not protesting the data gathering.[247] Differential privacy is viewed as the panacea to cure all data disclosure ailments when it is, in fact, a placebo.

It is also interesting that the entire responsibility of protecting the privacy of data is transferred to the public sector, while data brokers are free to sell any and all data. The process of disclosure is fraught with uncertainty since the malicious user can never be certain about the identity of the individual or the values of the variables for that record.[248] Purchasing data from the data broker is a better option for the malicious user when considering accuracy, effort, and cost, yet there is no protection against the sale of this data.[249] Indeed, the person who could be identified through reverse engineering has already identified himself in some other consensual manner, or data about that individual are already in the public domain.[250] Use of the privacy first assessment is guaranteed to prevent easy access to government data, but it fails to protect against a data broker's release of sensitive information.[251] The double standard contributes to data inequality and further negative rent seeking, which makes access to government data more difficult and renders the data held by data brokers and technology giants more valuable.

In an attempt to even the playing field, many commentators and scholars argue that data brokers, like the federal government, should be regulated by best practices known as fair information principles (FIPs).[252] Most privacy statutes incorporate the best practices for computer databases to ensure that a person who provides data for one purpose is not subjected to the use of that data for other purposes

---

245. *See id.*

246. Tom Simonite, *Apple's Privacy Pledge Complicates Its AI Push*, WIRED (July 14, 2017, 9:00 AM), https://www.wired.com/story/apple-ai-privacy/ [https://perma.cc/W3NY-UJ5Z].

247. Greenberg, *supra* note 238.

248. *See* Ohm, *supra* note 14, at 1710–11 (discussing the various proponents of anonymization of data prior to governmental release and the government's apparent acceptance of anonymization as the panacea for public release of data).

249. *Id.* at 1740. Ohm also challenges the theory that anonymization of data actually protects individual privacy. *Id.* at 1732.

250. *Id.* at 1725.

251. *Id.* at 1723.

252. *See, e.g.*, *id.* at 1733; Daniel J. Solove & Chris Jay Hoofnagle, *A Model Regime of Privacy Protection*, 2006 U. ILL. L. REV. 357, 358 (2006).

without his prior consent.[253] FIPs are a set of best practices for the collection, storage, and use of personal data by the government and the private sector.[254] The underlying philosophy of FIPs was the impetus behind the enactment of the Privacy Act.[255] However, scholars disagree as to whether, and to what extent, a mandatory application of FIPs for data brokers' activities would infringe upon their First Amendment rights.[256] A court examines the nature of the speech to determine whether a FIP or any other regulation on the distribution or receipt of data implicates First Amendment concerns.[257] Where data brokers direct communications to consumers encouraging them to buy more services or products, this form of speech is commercial in nature and subject to the *Central Hudson* analysis, which asks whether the activity is lawful and not misleading.[258] If so, then government may only restrict speech if "(1) it has a substantial state interest in regulating the speech, (2) the regulation directly and materially advances that interest, and (3) the regulation is no more extensive than necessary to serve the interest."[259] A state could arguably have a substantial interest in the collection, use, and reselling of data by commercial entities, and ensuring customers have the ability to opt out of such reuse may serve this interest.[260] Moreover, requiring the data broker

253. *See* Borgesius et al., *supra* note 20, at 2101–02.

254. Richards, *supra* note 61, at 1513.

255. *See id.* at 1510; discussion *supra* Part IV.B.

256. On one end of the spectrum, Eugene Volokh contends that most data privacy rules violate free speech, i.e. "violate my right to speak about you." *See* Eugene Volokh, *Freedom of Speech and Information Privacy: The Troubling Implications of a Right to Stop People from Speaking About You*, 52 STAN. L. REV. 1049, 1115–17 (2000). On the other end of the spectrum, Neil Richards advocates that FIPs "do not restrict the flow of data" but rather should be construed as confidentiality tools and that "data" are not entirely equal to "speech." *See* Richards, *supra* note 61, at 1512.

257. *See* U.S. West, Inc. v. FCC, 182 F.3d 1224, 1232–33 (10th Cir. 1999).

258. *See* Central Hudson Gas & Elec. Corp. v. Pub. Serv. Comm'n, 447 U.S. 557, 566 (1980).

259. *See* Revo v. Disciplinary Bd., 106 F.3d 929, 932 (10th Cir. 1997) (citing *Central Hudson*, 447 U.S. at 564–65); *see also U.S. West, Inc.*, 182 F.3d at 1235 (noting that while advancement of a privacy interest may be substantial, the government must articulate specifically how the regulation advances that interest, e.g., to avoid ridicule or harassment, rather than generically stating the restrictions are designed to protect privacy).

260. *See U.S. West, Inc.*, 182 F.3d at 1238–9 (noting that opt-out strategies from solicitations are "an obvious and substantially less restrictive alternative" to protect consumer privacy). Subsequent to the *U.S. West, Inc.* court's ruling against the FCC's opt-in regulations, the FCC modified them to apply in instances where consumer data are distributed to third parties rather than to the individual customer's carrier alone. *See* Customer Proprietary Network Information, 72 Fed. Reg. 31948, 31950 (June 8, 2007) (codified at 47 C.F.R. §§ 64.2001–.2011); *see also* Implementation of the Telecommunications Act of 1996, 17 FCC Rcd. 14860, 14875, 14883, 14889 (2002). Additionally, proponents of restricting the reuse of data gathered from data brokers have supported an initiative known as *Reclaim Your Name*. Julie Brill, Comm'r, Fed. Commc'n Comm'n, Keynote Address at the 23rd Computers Freedom and Privacy Conference: Reclaim Your

to make the underlying nature of its data available for corroboration promotes a substantial governmental interest by decreasing negative rent seeking.

VI. OPAQUENESS OF DATA BROKERS' DATA AND RESEARCH RESULTS

As previously noted, private entities have little restriction on where they get their data or how they share their data.[261] Based on intellectual property and trade secret protections, the release of their data is opaque, yet there are no consequences to this lack of transparency or lack of data protection.[262] At the same time, the government must protect the public data in its possession and must be transparent about its protection.[263] The fundamental discrimination between public and private sources of data, if not addressed, will lead to negative data inequality.

Due to a fundamental lack of transparency, it is unclear when, how, and what data are gathered by data brokers.[264] Data brokers gather much of the data without the specific knowledge or consent of the consumer.[265] Disconcertingly, a small number of sites receive the largest amount of traffic, meaning certain data aggregators and news sources control the majority of data consumers receive.[266] From a commercial perspective, approximately 81 percent of consumers

---

Name 10 (June 26, 2013), https://www.ftc.gov/sites/default/files/documents/public_statements/reclaim-your-name/130626computersfreedom.pdf [https://perma.cc/J9EW-QPGG].

261. *See* Ohm, *supra* note 184, at 1140–42.

262. *See* Mattioli, *supra* note 24, at 544–49.

263. *See* Joel R. Reidenberg, *The Transparent Citizen*, 47 LOY. U. CHI. L.J. 437, 439–40 (2015).

264. *See generally* 2014 DATA BROKER REPORT, *supra* note 12.

265. *Id.* at 46. Certainly, Internet users bear some responsibility to manage the data they share online, and this concept is contained within the Fair Data Practice Principles (FIPPs) (FIPs and FIPPs are used interchangeably throughout). *See* Solove, *supra* note 21, at 1882. Initially, these principles were designed to address part of the government's concern over the increase in digital data and including

> (1) transparency of record systems of personal data, (2) the right to notice about such record systems, (3) the right to prevent personal data from being used for new purposes without consent, (4) the right to correct or amend one's records, and (5) responsibilities on the holders of data to prevent its misuse.

Id. The underlying theme of FIPPs is a user's awareness that data are gathered and that the user consents to the gathering of the data. *Id.*

266. *See, e.g.*, Jeff Desjardins, *These Are the Top 100 Websites of the Internet, According to Web Traffic*, BUS. INSIDER, (Mar. 7, 2017, 8:08 PM), http://www.businessinsider.com/top-100-websites-web-traffic-2017-3.

conduct online research before making a purchase.[267] Forty-four percent of consumers commence their product search on Amazon's website, while 34 percent first consult Google, Bing, or Yahoo.[268] Regarding all searches, commercial or otherwise, website users search Google 100 billion times each month.[269] Not surprisingly, a top priority for marketers is how to improve their Internet presence, and 72 percent of marketers found that the most effective tool for their business has been ensuring their content's relevance to a consumer.[270] With the consumer visiting various sites—even though he reveals limited data on each—data can be aggregated by data brokers and compiled into a more detailed picture of the consumer and his private data.[271]

Many legal scholars criticize the inability to assess big data's pedigree as interfering with others' reuse of their data sets.[272] Traditional research methods define the research question, gather the data from a relevant data set, form a hypothesis, and test the hypothesis.[273] Once researchers publish traditional research, others test and challenge the research.[274] Modern commercial research alters this traditional method because big data often is considered proprietary in nature and not openly accessible for further analysis.[275] Accordingly, transparency of data is paramount to ensure accurate and thorough public research. Advancements in data brokers' techniques and algorithms are leading to more specific data, which are beneficial for

---

267. *See The Ultimate List of Marketing Statistics*, HUBSPOT, https://www.hubspot.com/marketing-statistics [https://perma.cc/8AQL-65Y5] (last visited Feb. 11, 2018).

268. *Id.*

269. *Id.*

270. *Id.* Video is increasingly a more popular tool for marketers, particularly through YouTube and Facebook. *Id.*

271. *See* Solove, *supra* note 21, at 1889.

272. *See, e.g.*, Mattioli, *supra* note 24, at 544–45.

273. *See* Tene & Polonetsky, *supra* note 68, at 354; *see also* Eszter Hargittai, *Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites*, 659 ANNALS AM. ACAD. POL. & SOC. SCI. 63, 73 (2015) (identifying common issues with the use of certain social media sites to conduct studies and noting females tend to use Twitter and Tumblr the most—while the less economically privileged do not—and African Americans use Twitter while Asian Americans were less likely to use LinkedIn); Andrew Moravcsik, *Tranparency: The Revolution in Qualitative Research*, 47 PS: POL. SCI. & POL. 48, 48 (2014) ("Transparency is the cornerstone of social science. Academic discourse rests on the obligation of scholars to reveal to their colleagues the data, theory, and methodology on which their conclusions rest. Unless other scholars can examine evidence, parse the analysis, and understand the processes by which evidence and theories were chosen, why should they trust—and thus expend the time and effort to scrutinize, critique, debate, or extend—existing research?").

274. *See* Tene & Polonetsky, *supra* note 68, at 355.

275. *Id.*; *see* Lev Manovich, *Trending: The Promises and the Challenges of Big Social Data*, MANOVICH (Apr. 28, 2011), http://manovich.net/content/04-projects/067-trending-the-promises-and-the-challenges-of-big-social-data/64-article-2011.pdf [https://perma.cc/YW3M-ZZXN].

marketing purposes but expose individuals to de-identification without their knowledge. For example, consumers are unaware that a grocery store can sell their purchasing data to third parties, and these third parties can then market to that consumer based on their grocery store purchase.[276] Moreover, website trackers can de-anonymize web browsing by linking to a person's Twitter and other social media accounts based on the person's clicking on a website link contained within the particular social media site.[277] In this regard, a person can be identified through the registration of his social media account and tied to the link, which is an indicator of interest in the content provided.[278]

Ideally, legislators could resolve this dilemma by requiring that data brokers provide access to their underlying data used in research. In those instances where the data broker does not wish to divulge the underlying data, they should be required to include a disclaimer noting the data are protected by trade secrets and not subject to independent review. As it is unlikely any such legislation would be enacted by the current administration, government or industry incentives should be considered as discussed below.

A. Opaque Data Can Lead to Erroneous Interpretations

In addition to the inability to challenge research based on data purchased from data brokers, use of the data can lead to erroneous conclusions. Research suggests there is a potential for incorporating errors and biases at every stage of the data and research process.[279] According to the FTC's 2014 Data Broker Report, some data brokers check the reliability of their data to ensure data are "internally consistent, corroborated by other sources, verifiable as legitimate, and that [they] encompass[] a sufficiently large portion of the population."[280] However, data brokers do not assess the accuracy of the government's

276. *See* Kuempel, *supra* note 32, at 219.

277. *See* Craig Mehall, *Study Finds Anonymous Browsing History Linkable to Individuals*, CQ ROLL CALL (Jan. 26, 2017), 2017 WL 370246.

278. *Id.*

279. For example, "social sorting involve[d] 'obtain[ing] personal and group data in order to classify people and populations according to varying criteria, to determine who should be targeted for special treatment, suspicion, eligibility, inclusion, access, and so on.'" *See* Borgesius et al., *supra* note 20, at 2092 (quoting David Lyon, *Surveillance as Social Sorting: Computer Codes and Mobile Bodies*, *in* SURVEILLANCE AS SOCIAL SORTING: PRIVACY, RISK AND DIGITAL DISCRIMINATION 13 (David Lyon ed., 2002)).

280. *See* 2014 DATA BROKER REPORT, *supra* note 12, at 16.

data or other publicly available data that they gather before incorporating them into their analysis.[281]

In this regard, the choice of the data set used to make predictions, defining the problem to be addressed through big data, and the decisions based on the results of big data analysis could lead to potential discriminatory harms, which are examined below through the following examples: (1) the advent of fake news and the public's belief in such news, (2) the effect of inaccurate background checks, and (3) the unintended consequences of misinterpreting data.[282] Other researchers have noted that these concerns are overstated—or are simply not new—and emphasize that rather than disadvantaging minorities, big data can create opportunities for low-income and underserved populations because the data can identify discrepancies or previously unknown needs.[283] However, it is becoming increasingly apparent that data can be manipulated either intentionally or unintentionally.[284] One element that both public and nonpublic data have in common is the effect human judgment can have on the accumulation and assessment of the data.[285] The outcome of the analysis is dictated by the type of data collected, the question presented, the pool of subjects in the dataset, the method of collection, and its assessment. The method of culling and trimming data is known as "cleaning" the data.[286] The process is highly subjective, and the same data analysis could lead to different results depending on the person or persons conducting the analysis.[287] Because of this highly subjective method of research, proponents for data transparency in research are growing.[288]

1. Fake News

A significant example of the need to ensure data's accuracy can be seen in the aftermath of the 2016 elections and the idea of "fake news." Interestingly, people tend to believe what they read.[289] The fake news that dominated Facebook preceding the election was created by

---

281. *Id.*

282. *See* Tene & Polonetsky, *supra* note 68, at 353–54.

283. *Id.* at 355–56; *see* RAMIREZ ET AL., *supra* note 62, at 2–3 (detailing data gathered through public workshops regarding big data).

284. *See* Tene & Polonetsky, *supra* note 68, at 353–54.

285. *See* Mattioli, *supra* note 24, at 546.

286. *Id.* at 561.

287. *Id.*

288. *Id.*

289. *See* Interview by Dave Davies with Craig Silverman, Editor, BuzzFeed News (Dec. 14, 2016), www.npr.org/2016/12/14/505547295/fake-news-expert-on-how-false-stories-spread-and-why-people-believe-them [https://perma.cc/N2H4-8GDY] (interviewing Craig Silverman of BuzzFeed News, who has spent years studying media inaccuracy).

teenagers in Macedonia to make a profit from the pro-Trump movement.[290] In one instance, fake news circulating on Facebook that Hillary Clinton and the owner of a Washington, DC, pizzeria ran a child prostitution ring out of the restaurant provided the impetus behind a gunman's attempt to kill the restaurant owner.[291] Even though the story appeared outlandish, people—including the perpetrator—believed the fake news.[292]

Despite the numerous discussions of fake news, 84 percent of the US population feels at least somewhat confident about spotting fake news, 39 percent of whom feel very confident that they can spot fake news and 45 percent of whom feel somewhat confident.[293] However, an Ipsos poll conducted for Buzzfeed News found 75 percent of Americans believed the fake news stories they had heard from the election.[294] During the election, Facebook had altered its algorithms that prioritized what its users saw by decreasing news media feeds and increasing posts and updates from friends and families.[295] Because Facebook owns the data, only it can say whether this helped the proliferation of fake news and whether its current efforts have reduced the circulation of fake news.[296] Approximately 1.8 billion monthly users and nearly half of all US adults read Facebook news.[297] Facebook has

---

290. *See id.*

291. Brett Edkins, *Americans Believe They Can Detect Fake News. Studies Show They Can't.*, FORBES (Dec. 20, 2016, 1:46 PM), https://www.forbes.com/sites/brettedkins/2016/12/20/americans-believe-they-can-detect-fake-news-studies-show-they-cant/#5c1126fd4022 [https://perma.cc/7CVN-7NPR].

292. *See* German Lopez, *Pizzagate, the Fake News Conspiracy Theory That Led a Gunman to DC's Comet Ping Pong, Explained*, VOX (Dec. 8, 2016, 11:15 AM), https://www.vox.com/policy-and-politics/2016/12/5/13842258/pizzagate-comet-ping-pong-fake-news [https://perma.cc/N5B4-KVBT].

293. *See* Michael Barthel, Amy Mitchell & Jesse Holcomb, *Many Americans Believe Fake News Is Sowing Confusion*, PEW RES. CTR. (Dec. 15, 2016), http://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/ [https://perma.cc/2DUH-LXXQ].

294. Edkins, *supra* note 291 (noting that 84 percent of respondents believed the fake news story that "Donald Trump Sent His Own Plane to Transport 200 Stranded Marines" and that three-fourths of Trump supporters believed the fake news story that Pope Francis endorsed Donald Trump). A Stanford University study supports the findings that Americans rarely identify fake stories as false. *Id.*

295. Mike Isaac & Sydney Ember, *Facebook to Change News Feed to Focus on Friends and Family*, N.Y. TIMES (June 29, 2016), https://www.nytimes.com/2016/06/30/technology/facebook-to-change-news-feed-to-focus-on-friends-and-family.html [https://perma.cc/6XJ3-YNBR].

296. Editorial, *Facebook and the Digital Virus Called Fake News*, N.Y. TIMES (Nov. 19, 2016), https://www.nytimes.com/2016/11/20/opinion/sunday/facebook-and-the-digital-virus-called-fake-news.html [https://perma.cc/D5U5-2NB2].

297. Nick Wingfield, Mike Isaac & Katie Benner, *Google and Facebook to Take Aim at Fake News Sites*, N.Y. TIMES (Nov. 14, 2016), https://www.nytimes.com/2016/11/15/technology/google-will-ban-websites-that-host-fake-news-from-using-its-ad-service.html

agreed to partner with third parties to flag fake news articles and alert users before they share the false news.[298] Both Google and Facebook, meanwhile, collaborated with journalists to curb fake news stories in advance of the 2017 French elections.[299] They carried out these efforts through trending and data mining techniques to detect problematic stories, provide cross-checking resources for readers, and attach warning labels to suspect stories.[300]

While fake news is not a privacy issue, it has one important lesson: in the absence of transparent data, research will be unreliable, and researchers will not be incentivized to provide quality or accurate research as long as others will not be able to easily cross-check their research.[301] Researchers and those funding the research are responsible for ensuring transparency.[302] For example, the federal government often requires federally funded research to be made public within a certain timeframe, journal editors require adherence to certain publishing guidelines, and some journal editors provide meaningful consequences for research misconduct.[303] However, these rules do not apply to nonfederally funded research. In addition to the obvious

---

[https://perma.cc/K5MF-RHGY]. The *New York Times* calls the republishing of fake news a "digital virus." Editorial, *supra* note 296.

298. Ivana Kottasova, *Facebook, Google to Help Fight Fake News Ahead of French Elections*, CNN MONEY (Feb. 6, 2017, 9:40 AM), http://money.cnn.com/2017/02/06/technology/france-elections-fake-news-facebook-google/index.html [https://perma.cc/A5RE-XQ86]. Facebook will post "disputed by 3rd Party Fact-Checkers" beneath the fake stories, will send an alert when the story is shared, and will rank the disputed stories lower in the news feed and prevent them from transforming into promotions. *See* Public Voice, *Facebook to Start Putting Red Label on 'Fake News'*, PUBLICVOICE (Dec. 15, 2016), http://www.publicvoice.ng/newsarticles/show_detail/74037/Facebook_to_start_putting_red_label_on_fake_news#sthash.m59YUOkD.dpbs [https://perma.cc/FPY6-EAX5]. Likewise, Google intends to prohibit websites from selling fake news. Alex Hern, *Facebook and Google Move to Kick Fake News Sites off Their Ad Networks*, GUARDIAN (Nov. 15, 2016, 7:07 AM), https://www.theguardian.com/technology/2016/nov/15/facebook-google-fake-news-sites-ad-networks [https://perma.cc/TNL7-PP9N].

299. Kottasova, *supra* note 298.

300. Natasha Lomas, *Google and Facebook Partner for Anti-Fake News Drive During French Election*, TECHCRUNCH (Feb. 6, 2017), https://techcrunch.com/2017/02/06/google-and-facebook-partner-for-anti-fake-news-drive-during-french-election/ [https://perma.cc/JM6M-W9G8].

301. *See* Moravcsik, *supra* note 273, at 50. According to the American Political Science Association, transparency in research has three distinct characteristics: (1) data transparency, which provides the reader with the evidence used to support the claims; (2) analytic transparency, the "process by which an author infers that evidence supports a specific descriptive, interpretive, or causal claim"; and (3) production transparency, which provides the reader with the facts surrounding the reason the author chose a particular source for his research. *Id.* at 48–49.

302. Patricia K. Baskin, *Transparency in Research and Reporting: Expanding the Effort Through New Tools for Authors and Editors*, EDITAGE (July 20, 2015), http://www.editage.com/insights/transparency-in-research-and-reporting-expanding-the-effort-through-new-tools-for-authors-and-editors [https://perma.cc/EYC4-DKYA].

303. *Id.*

potential inaccuracy issues, the use of big data poses potential ethical problems for society.[304]

2. Inaccurate Information and Credit Reports

According to one of the leading and largest data brokers, 30 percent of data brokers' data are inaccurate.[305] One expensive example was detected by the FTC when Spokeo, a data broker, marketed an employment screening tool with inaccurate profiles.[306] In addition to an $800,000 fine, one affected consumer sued Spokeo under the Fair Credit Reporting Act (FCRA) for the publication of inaccurate data about his personal and employment background, believing it would harm his future employment possibilities.[307] In May 2016, the Supreme Court remanded the matter to the Ninth Circuit to determine whether the consumer had alleged a concrete and particularized injury from Spokeo's violation of the FCRA but implicitly recognized there might be instances where such an injury automatically exists.[308] The FCRA applies to "consumer reporting agencies" that compile data into "credit reports" which are then used to score an individual's

304. *See* Danah Boyd & Jacob Metcalf, Example "Big Data" Research Controversies 1 (Nov. 10, 2014) (unpublished report), http://bdes.datasociety.net/wp-content/uploads/2016/10/ExampleControversies.pdf [https://perma.cc/87RS-NC5R] (identifying the ethical concerns between data supplied for certain surveys or commercial purposes being re-tooled and utilized for other purposes, such as governmental policy decisions); *see also* Mattioli, *supra* note 24, at 561 (citing a cancer research institution where data collectors used available data such as height or weight to determine the gender of transsexual patients because the healthcare institution had routinely labeled such patients as having an "unknown" gender).

305. *See* Lipman, *supra* note 113, at 782; *see also* UPTURN, CIVIL RIGHTS, BIG DATA, AND OUR ALGORITHMIC FUTURE 13 (2014) [hereinafter SEPTEMBER 2014 REPORT], https://bigdata.fairness.io/wp-content/uploads/2015/04/2015-04-20-Civil-Rights-Big-Data-and-Our-Algorithmic-Future-v1.2.pdf [https://perma.cc/BJA6-CRB5] (detailing racial bias issues in the governmental work-eligibility of potential employees, which carries a 20 percent higher error rate for those who are foreign born as opposed to those born in the United States); Bobby Allyn, *How the Careless Errors of Credit Reporting Agencies Are Ruining People's Lives*, WASH. POST (Sept. 8, 2016), https://www.washingtonpost.com/posteverything/wp/2016/09/08/how-the-careless-errors-of-credit-reporting-agencies-are-ruining-peoples-lives/?utm_term=.a8083de9383b [https://perma.cc/MR9L-QLU8].

306. *See* Press Release, Fed. Trade Comm'n, Spokeo to Pay $800,000 to Settle FTC Charges Company Allegedly Marketed Data to Employers and Recruiters in Violation of FCRA (June 12, 2012) [hereinafter Press Release, Spokeo to Pay $800,000], https://www.ftc.gov/news-events/press-releases/2012/06/spokeo-pay-800000-settle-ftc-charges-company-allegedly-marketed [https://perma.cc/VC3A-XDSU].

307. Chris Morran, *Why You Should Care About This Lawsuit Against a Data Company You've Probably Never Heard Of*, CONSUMERIST (Aug. 15, 2017, 3:50 PM), https://consumerist.com/2017/08/15/why-you-should-care-about-this-lawsuit-against-a-data-company-youve-probably-never-heard-of/ [https://perma.cc/MG82-D4LG].

308. *See* Spokeo, Inc. v. Robins, 136 S. Ct. 1540, 1545 (2016).

creditworthiness.[309] Credit reports contain "any information . . . bearing on a customer's credit worthiness, credit standing, credit capacity, character, general reputation, personal characteristics, or mode of living."[310] Data must be "used or expected to be used or collected" to serve as "a factor in establishing the consumer's eligibility for" credit, insurance, or employment for it to be subject to the FCRA.[311] If data are subject to the FCRA, the reporting agency must comply with a variety of obligations surrounding the collection, use, and right to challenge the data.[312] The trouble with this in the context of big data aggregation is that data mining can often be inaccurate and provide skewed research results, culminating in a lender's refusal to issue a loan or an employer's decision to withhold employment.[313] Data brokers must be vigilant in the use of the data they gather and sell to avoid intentional discrimination claims. More importantly, access to transparent data is paramount so that researchers can ferret out instances of intentional discrimination. The Authors assert the continued failures to do so will lead to data inequality and negative rent seeking.

3. Misinterpretation of Data

Biases can occur at any stage in this modern form of research, including the collection and the analysis stages.[314] If one considers the assessment of big data once the anonymization process is employed, one can see how easily research results could be faulty based on the important need for individual privacy.[315] Researchers identified an exemplification of an unintentionally erroneous analysis in 2012 involving Hurricane Sandy and the twenty million tweets and data

309. *See id.*; Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Prediction*, 89 WASH. L. REV. 1, 17 (2014) (discussing the FCRA).

310. 15 U.S.C. § 1681a(d)(1) (2012).

311. *Id.* § 1681a(d)(1)(A).

312. *See Spokeo*, 136 S. Ct. at 1545.

313. *See* Mikella Jurley & Julius Adebayo, *Credit Scoring in the Era of Big Data*, 18 YALE J.L. & TECH. 148, 153 (2016); *see also* SEPTEMBER 2014 REPORT, *supra* note 305, at 13; Allyn, *supra* note 305.

314. *See* Yoni Har Carmel & Tammy Harel Ben-Shahar, *Reshaping Ability Grouping Through Big Data*, 20 VAND. J. ENT. & TECH. L. 87, 112–14 (2017); Kate Crawford, *The Hidden Biases in Big Data*, HARV. BUS. REV. (Apr. 1, 2013), https://hbr.org/2013/04/the-hidden-biases-in-big-data [https://perma.cc/57U8-2B5A].

315. The simplest form of de-identification is the removal of all identifying data such as names, addresses and any other personally identifiable data. *See* Mattioli, *supra* note 24, at 566. However, doing so can weaken the effectiveness of the data. *Id.* More sophisticated methods are also used whereby fake values are inserted into the mix of data, secreting the identity or other data of the population. *See id.* Data analytics assist in mining data including air-sensor feedback and weather data. *See* Breggin & Amsalem, *supra* note 18, at 10985–86.

from FourSquare between October 27 and November 1.[316] Some of the data reflected grocery purchases that increased the evening before the hurricane and nightlife spending the night after the hurricane.[317] However, more surprisingly, the majority of the tweets came from Manhattan, creating an impression that the hurricane significantly impacted Manhattan when in fact it had not.[318] The locations most affected by the hurricane had minimal amounts of Twitter messages.[319] Accordingly, if researchers had not thoroughly analyzed the data, the results could have been improperly skewed.[320]

Another interesting example of a potentially erroneous analysis of big data was Boston's attempt to detect and rectify its pothole problems through an application called StreetBump, which passively detects potholes through GPS and acceleration data.[321] The issue with this data collection method was that it did not account for individuals in locations with limited cell phone use or limited access to cars, occurring in generally lower income areas.[322] Fortunately, the individuals associated with collecting the data recognized this as a possibility and accounted for the discrepancies.[323] These examples demonstrate that without alternate, open sources of verifiable data, erroneous predictions and results are more likely to occur.

4. Self-Regulatory Attempts to Rectify Misinformation

Notably, one data broker—Acxiom—recognizes the potential pitfalls of erroneous data collection or interpretation is attempting to combat the issues through a website that allows individuals to log in and correct any errors to their information.[324] It also provides an option for consumers to opt out of data collection.[325] However, some critics argue that the opt-out mechanism is simply another means by which the data broker giant accesses more information about individuals

316. *See* Crawford, *supra* note 314.

317. *Id.*

318. *Id.*

319. *Id.*

320. *Id.*

321. *Id.*

322. *Id.*

323. *Id.* However, Kate Crawford's article also noted Google's failure to accurately predict flu trends and its failure to broadcast why its data was skewed.

324. *See Consumer Data Information*, ACXIOM, https://www.acxiom.com/about-us/privacy/consumer-data-information/ [https://perma.cc/PRM5-GBC5] (last visited Feb. 11, 2018).

325. *See Acxiom Corporation's Online Opt-Out*, AXCIOM, https://isapps.acxiom.com/optout/optout.aspx [https://perma.cc/G4M6-ZP4Z] (last visited Feb. 11, 2018).

instead of truly allowing them to opt out.[326] In light of the data broker's attempt at transparency and opportunities for consumers to opt out of the collection of data, there could be some interest in self-regulation by these entities. This might include advanced marketing techniques and trademarks signifying the entities' commitment to transparency and cooperation similar to the "Fair Trade" marketing movement for consumer products.[327]

B. Intentional Manipulation of Big Data

In the context of research, even more troubling than unintentionally inaccurate data results are those results that have been intentionally manipulated by private sources such as (1) the Facebook emotion experiment, (2) the Facebook voting experiment, and (3) the OkCupid compatibility experiment, each discussed below. Big data's influence on public opinion has become increasingly concerning. Joe Turow, Cass Sunstein, and other researchers have detailed how commercial entities—like Facebook—can modify their consumers' opinions simply through the types of data such entities allow their users to see.[328] Potential voters can be recipients of targeted election data based on their fears, interests, or supported causes identified through their online usage.[329] Unlike public researchers who are governed by federal regulations, private researchers are subject to limited privacy restrictions and their own privacy policies.[330] The Federal Policy for the Protection of Human Research Subjects, known as the "Common Rule," governs human subject research conducted by public researchers (those that receive federal funding or those who voluntarily comply with its

326. *See* Will Simonds, *Acxiom's Letting You See the Data They Have About You (Kind Of)*, ONLINE PRIVACY BLOG (Sept. 4, 2013), https://www.abine.com/blog/2013/acxioms-letting-you-see-data/ [https://perma.cc/CBT6-FBXT]. Consumer advocacy group Stop Data Mining has compiled a master list of how consumers can opt out of data collection. *See Opt Out List*, STOPDATAMINING, http://www.stopdatamining.me/opt-out-list/ [https://perma.cc/9CF3-ZD95] (last visited Feb. 11, 2018).

327. For example, the nonprofit organization Fair Trade USA certifies those products that meet required minimum standards for fair prices, wages, working conditions, and environmental and community-based protections. *Who We Are*, FAIR TRADE CERTIFIED, https://www.fairtradecertified.org/about-us [https://perma.cc/6CXF-R9FX] (last visited Feb. 6, 2018).

328. *See* Tene & Polonetsky, *supra* note 68, at 359.

329. *Id.* at 360 (recognizing that the federal government is using big data to assist in its policy analysis); *see* Breggin & Amsalem, *supra* note 18, at 10991. An interesting example includes data that shows a significant draw of power can help law enforcement locate marijuana use. Breggin & Amsalem, *supra* note 18, at 10992.

330. *See* James Grimmelmann, *The Law and Ethics of Experiments on Social Media Users*, 13 COLO. TECH. L.J. 219, 221 (2015).

terms for privately funded research).[331] The Common Rule's fundamental policy is that researchers fully and completely inform the subjects regarding the data gathered about them and such data's use.[332]

A prime example of intentional manipulation of research is Facebook's emotion research where researchers and Facebook exposed certain subscribers to negative newsfeeds and other subscribers to positive newsfeeds to determine whether the groups exposed to more negative feeds had more negative postings.[333] Those with the more negative posts demonstrated more negative reposts to others, and those receiving the positive posts reposted more positive material.[334] Both Facebook and the academic researchers faced significant criticism for what the public perceived was unethical human research.[335] The criticism included concerns over the subject pool, as it had no age filter and could have included minors without parental informed consent and included nationals from other countries, potentially violating international data protection laws.[336] Additionally, the researchers did not follow the Common Rule protocols.[337] Common Rule protocols provide parameters surrounding human research studies and apply to certain categories of government-funded research, but since Facebook is a private entity, such regulations were inapplicable.[338]

In another example of intentional manipulation of data, researchers hired by Facebook experimented on approximately

---

331. *Id.*

332. *Id.* at 227.

333. *See* Jane R. Bambauer, *All Life Is an Experiment. (Sometimes It Is a Controlled Experiment.)*, 47 LOY. U. CHI. L.J. 487, 489 (2015).

334. *Id.*

335. *See* Calli Schroeder, Note, *Why Can't We Be Friends? A Proposal for Universal Ethical Standards in Human Subject Research*, 14 COLO. TECH. L.J. 409, 418 (2016). The study sparked controversy regarding ethical standards in research and whether researchers (particularly those working for public universities) should comply with the Common Rule for these types of research projects despite the project being privately funded. *Id.* Seven hundred thousand unwitting Facebook subscribers were the subject of the experiment conducted between Facebook and researchers from Cornell University. Bambauer, *supra* note 333, at 489.

336. Schroeder, *supra* note 335, at 412–13.

337. The Common Rule requires all federally funded entities or research to follow certain institutional review board standards that examine the risks, minimize those risks, identify the benefits, compare the reasonableness of the risks to the benefits, ensure subjects are fully informed and consent, and provide periodic review and monitoring. *Id.* at 412–13.

338. The regulations require the participants to have specific knowledge and consent. *Id.* Although Facebook contends its privacy policy covered the informed consent, it was not until four months after the study that Facebook revised its policy to state "that user data could be used for several internal operations purposes, including 'troubleshooting, data analysis, testing, research and service improvement.'" *Id.* at 423–24 (citations omitted). Moreover, the study included adolescents and minors, raising red flags concerning mood manipulation and lack of parental knowledge and consent. *Id.* at 412.

sixty-one million Facebook users immediately preceding the 2010 midterm elections.[339] The experiment divided users into two groups: in one, the researchers showed users a "go vote" message in a plain box; the other group was shown the same box with the addition of thumbnail pictures of their friends who had clicked on "I voted."[340] After the election, the researchers compared the two groups through voting poll records and determined the latter had hundreds of thousands of voters whereas the former group did not.[341] How Facebook conducted this research, and the extent of its users' knowledge of their participation in the experiment, is solely within Facebook's control and exemplifies the need for disclosure obligations. This particular Facebook experiment exemplifies the need for disclosure obligations for many reasons, including, *inter alia*, to combat fake news, to combat data use without meaningful consent,[342] and to avoid undemocratic election tampering.[343]

---

339. *See* Zeynep Tufekci, Opinion, *Mark Zuckerberg Is in Denial*, N.Y. TIMES (Nov. 15, 2016), www.nytimes.com/2016/11/15/opinion/mark-zuckerberg-is-in-denial.html [https://perma.cc/YY2F-LA2E]. Facebook users over the age of eighteen were automatically included in the experiment. Robert M. Bond et al., *A 61-Million-Person Experiment in Social Influence and Political Mobilization*, 489 NATURE 295, 295 (2012), https://www.nature.com/nature/journal/v489/n7415/full/nature11421.html [https://perma.cc/E7PW-M2EZ].

340. *See* Bond et al., *supra* note 339, at 295.

341. *See id.* at 297.

342. Despite a research policy of "informed consent," the joint university and Facebook experiment lacked any "informed consent" when it "automatically included" all Facebook users. *See Frequently Asked Questions*, U.C. SAN DIEGO, https://irb.ucsd.edu/FAQs.FWx [https://perma.cc/3EJZ-XS56] (last visited Feb. 11, 2018).

343. An even more serious question is the extent to which such experiments can be used to subvert democracy. The authors of the study say that a conservative estimate of the increase in turnout was 340,000 and speculate that it might have been as high as 1.4 million. *See* Bond et al., *supra* note 339, at 295; *see also* Siva Vaidhyanathan, Opinion, *Facebook Wins, Democracy Loses*, N.Y. TIMES (Sept. 8, 2017), https://www.nytimes.com/2017/09/08/opinion/facebook-wins-democracy-loses.html?_r=0 [https://perma.cc/GV3V-2MYR]. While increasing participation in elections may be a noble cause, it could be easily used for ignoble purposes. Just as Facebook messages can encourage voting across the political spectrum, it is possible that it could also be used to encourage voting for one candidate and discourage voting for another. There is some evidence to support that such attempts occurred during the 2016 presidential election. *Id.* Given that Hillary Clinton was likely to have lost the electoral votes in Michigan, Pennsylvania, and Wisconsin by less than 80,000 votes in total, Facebook's ability to increase turnout by over 340,000 votes in a non-presidential year election may be viewed with alarm. *See* Philip Bump, *Donald Trump Will Be President Thanks to 80,000 People in Three States*, WASH. POST (Dec. 1, 2016), https://www.washingtonpost.com/news/the-fix/wp/2016/12/01/donald-trump-will-be-president-thanks-to-80000-people-in-three-states/?utm_term=.47b8c9a95ac2 [https://perma.cc/K8RF-NG2E]. And while news that Mark Zuckerberg is running for president may be pure speculation, questions about Facebook's ability to influence elections takes on added significance. Abby Ohlheiser, *Even Mark Zuckerberg Can't Stop the Meme That He Is Running for President*, WASH. POST (Aug. 3, 2017), https://www.washingtonpost.com/news/the-intersect/wp/2017/08/03/even-mark-zuckerberg-cant-stop-the-meme-that-he-is-running-for-president/?utm_term=.2cb2ca977bb1 [https://perma.cc/26A5-NLHE].

A growing concern among privacy advocates is the advent of psychological targeting through big data.[344] Data brokers know "your age, income, favorite cereal and when you last voted."[345] Companies or politicians can target their marketing efforts to correspond with an individual's psychological profile; for instance, if a person is deemed a worrier, data brokers may show that person ads or news about the dangers of the Islamic State to assist in driving him toward a political candidate or product.[346] This psychological profiling is also known as "emotion analysis," and social media sites conduct this exact form of profiling.[347] Companies engaged in psychological profiling note that the United States is an easy target, as its privacy laws surrounding the data gathered on individuals are minimal compared to the European Union's privacy laws.[348]

For example, OkCupid conducted psychological profiling experiments on subscribers to understand which aspects of their profile were the most relevant.[349] Unbeknownst to its subscribers, OkCupid experimented on five hundred users, telling a group they were not compatible with one another and another group that they *were* compatible with one another.[350] The experiment noted that when individuals are told they are compatible, they act as if they are, even when they are not.[351] Likewise, those who believe they are incompatible do not seek further contact with each other despite their actual compatibility.[352] Again, OkCupid is a private entity and merely relies on its terms of use to conduct frequent and ongoing research on its subscribers.[353]

---

344. *See* Nicholas Confessore & Danny Hakim, *Data Firm Says 'Secret Sauce' Aided Trump; Many Scoff*, N.Y. TIMES (Mar. 6, 2017), https://www.nytimes.com/2017/03/06/us/politics/cambridge-analytica.html?_r=1 [https://perma.cc/Y5RB-LKWN].

345. *Id.*

346. *Id.*

347. *Id.*

348. *Id.*; *see also* Hannah L. Cook, *Flagging the Middle Ground of the Right to Be Forgotten: Combatting Old News with Search Engine Flags*, 20 VAND. J. ENT. & TECH. L. 1, 9 (2017) (remarking this difference connotes "a clash between a US conception of privacy as a property right that can be bargained away and a European view that privacy encompasses a human dignity that cannot be exchanged or removed").

349. *See* Grimmelmann, *supra* note 330, at 223–24.

350. *Id.* at 224.

351. *Id.*

352. *Id.*

353. "[I]f you use the Internet, you're the subject of hundreds of experiments at any given time, on every site," Christian Rudder, President of OKCupid, wrote on the company's blog, "That's how websites work." Molly Wood, *OKCupid Plays with Love in User Experiments*, N.Y. TIMES (July 28, 2014) (quoting Christian Rudder, *We Experiment on Human Beings! (So Does Everyone Else)*, OKCUPID BLOG (July 27, 2014), https://theblog.okcupid.com/we-experiment-on-human-beings-

Using the Internet today risks the disclosure of a user's deepest and most intimate secrets. In a society dependent on the Internet for much of its daily interactions in both the social and business context, the population is in the unenviable position of "caveat user."[354] Nonetheless, the user is unwittingly oblivious of the full ramifications of his Internet presence. Data brokers' experiments reflect a common theme: data brokers and other entities who utilize the public Internet can and do conduct research on individual users. Whether designed significantly to impact society or otherwise, the effects are the same. Those with access to private data therefore control the future of research and contribute to data inequality.[355]

VII. IMPACT ON RESEARCH AND THE NEED FOR A CROSS-POLLINATION[356] OF DATA BETWEEN DATA BROKERS AND THE GOVERNMENT

Government agencies protect an individual's privacy through data anonymization and refusal to release discretionary data.[357] For the researcher without access (or resources) to data from brokers, the primary source of research data remains through government agencies' release of data.[358] Data released by government agencies also serve an

5dd9fe280cd5 [https://perma.cc/6UV7-5686]), https://www.nytimes.com/2014/07/29/technology/okcupid-publishes-findings-of-user-experiments.html?_r=0 [https://perma.cc/6JNF-2KMD].

354. *See Caveat User: Data Mining and Sneaky Services Providers*, CHRISTIANSEN IT L., http://christiansenlaw.net/2011/10/caveat-user-data-mining-and-sneaky-services-providers/ [https://perma.cc/YV59-UF4H] (last visited Feb. 12, 2018) ("If you have notice of and an opportunity to read the terms of use and choose not to do so[,]. . . you're still bound by the information they contain and the agreements they include.").

355. While some results and analysis may be unintentionally erroneous, researchers and legal scholars have noted big data research can be used to intentionally discriminate against certain populations, and companies can segregate data by certain zip codes, disparately impacting lower income individuals. *See* Scott R. Peppet, *Regulating the Internet of Things: First Steps Toward Managing Discrimination, Privacy, Security, and Consent*, 93 TEX. L. REV. 85, 117–18, 120 (2014) (arguing because data can be gathered from a multitude of devices, vendors and others can take action based on this data to the consumer's detriment). Once the government releases the data, data brokers often collect the data, analyze it, and resell it for a variety of purposes including marketing, credit scoring, or screening job applicants. *See* Borgesius et al., *supra* note 20, at 2092 (noting that household data like whether a smoker lives in the home could be used to decline insurance). Banks and other entities can use big data to assist them in determining the credit worthiness of potential clients. *See* Citron & Pasquale, *supra* note 309, at 17 (discussing the FCRA).

356. A cross-pollination of data would necessarily require cooperation between the public and private sector. The Authors contend where information is gathered in one, the other should have the right to access it for certain purposes free of charge.

357. *See* Altman et al., *supra* note 74, at 1991.

358. *See id.* at 1970.

important societal purpose by providing information to the public.[359] The data released by the government are precisely the type of data that data brokers sell.[360] Without the governmental release of data, or with the release of data so anonymized as to make the data useless, only those researchers with adequate funding or those with relationships with data brokers will constitute the field of future research. Thus, data inequality will lead to further income inequality. Without access to unbiased, verifiable data, the public will not know whether the product, news article, or scientific conclusion derives from accurate data or whether that data has been intentionally or unintentionally manipulated by data gathered and supplied by the opaque big business of data brokers. If the public is told what to think and what is accurate without being able to properly and accurately challenge that data, the public can be further divided between the educated, powerful rich and the manipulated, weakened poor, contributing to negative rent seeking.

To combat data inequality and negative rent seeking, the Authors recommend several potential solutions: (1) legislation requiring data brokers and other online services to provide full and complete disclosure to users regarding the information they collect, its reuse, and potential aggregation and resale; (2) legislation allowing for users to opt out of data collection and reuse without forgoing use of the data broker's or other online provider's services; (3) legislation requiring data brokers to share underlying data with the government and researchers necessary for research in the fields of public welfare and national security; (4) modification of the government's analysis in discretionary release of information to include evaluating a person's revelation of the same information sought to be released to data brokers or other online providers; (5) encouraging the data broker industry to voluntarily provide disclosure and opt-out mechanisms for users; and (6) encouraging the data broker industry to voluntarily adopt a certification of transparency that could be used as a marketing tool while simultaneously reducing the negative consequences associated with their opaque data process.

The Authors recognize two significant obstacles to the legislative suggestions. First, the Trump administration has evidenced an intent to eliminate and reduce government regulations and has reduced both the FTC's and FCC's authority to regulate the data broker business.[361] Thus, it is highly unlikely any direct legislation or regulation in this

359. *Id.*

360. *Id.* at 1987–88.

361. *See* Cecilia Kang, *Congress Moves to Strike Internet Privacy Rules from Obama Era*, N.Y. TIMES (Mar. 23, 2017), https://www.nytimes.com/2017/03/23/technology/congress-moves-to-strike-internet-privacy-rules-from-obama-era.html?_r=0 [https://perma.cc/BD8L-PY3P].

area will be forthcoming in the near future. Second, any legislative disclosure or disclaimer obligations must meet the exacting standards of *Citizens United* and *Sorrell*.[362] Regulating data brokers and the data they collect has been met with numerous legal and scholarly challenges on grounds ranging from intellectual property rights to contract rights and constitutional rights.[363] Individuals currently do not have copyright protection for the facts they release to the Internet.[364] Further, there is no fundamental right to privacy in most consumer activity on the Internet, and the consumer often relinquishes what privacy does exist through consent to the terms of use and service drafted by the Internet provider.[365] Some argue that a movement similar to the European Union's is instructive and that there is a fundamental right to specific knowledge about what data brokers gather and how others use it, with an affirmative right to opt out.[366] This issue becomes more worrisome when considering the ease of purchasing data from data brokers, which results in an increased demand for privacy in the release of government data.[367] The idea that data brokers must share their underlying data with the government, even for limited topics, likely would face extreme opposition.[368]

---

362. Sorrell v. IMS Health Inc., 564 U.S. 552, 571–72 (2011); Citizens United v. Fed. Election Comm'n, 558 U.S. 310, 371 (2010).

363. *Compare* FRED H. CATE, PRIVACY IN THE INFORMATION AGE 30 (1997) (arguing privacy is "an antisocial construct . . . [that] conflicts with other important values within the society, such as society's interest in facilitating free expression"), *with* Jane Bambauer, *Is Data Speech?*, 66 STAN. L. REV. 57, 57 (2014), *and* David Post, *Cyberprivacy, or What I (Still) Don't Get*, 20 TEMP. POL. & C.R.L. REV. 249, 251 (2011) ("[O]ne person's privacy is very often another person's infringement of the freedom to speak."), *and* Neil M. Richards, *Intellectual Privacy*, 87 TEX. L. REV. 387, 390 (2008) ("Indeed, when it comes to database regulation, many feel that any government regulation of private information flows raises serious First Amendment issues.").

364. *See, e.g.*, Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., 499 U.S. 340, 344 (1991) (holding facts such as telephone numbers are generally not copyrightable); Peter K. Yu, *The Political Economy of Data Protection*, 84 CHI-KENT L. REV. 777, 780–81 (2010) (discussing implications of the *Feist* holding on database protections).

365. *See* Diana Liebenau, Note, *What Intellectual Property Can Learn from Privacy, and Vice Versa*, 30 HARV. J.L. & TECH 285, 296 (2016) ("On social media, users routinely grant a worldwide, non-exclusive, royalty-free license . . . to their copyrighted content by accepting the Terms of Service agreements."). Users often do not understand what they are agreeing to and, thus, their voluntariness is questionable. *See* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 667 (2016).

366. *See, e.g.*, Bradyn Fairclough, Note, *Privacy Piracy: The Shortcomings of the United States' Data Privacy Regime and How to Fix It*, 42 J. CORP. L. 461, 478–80 (2016) (arguing FIPs' obligations reflect an acknowledgment that there is a "right" to personal data and discussing the EU's treatment of data protection as a fundamental human right).

367. *See* Dwork, *supra* note 237, at 91.

368. It is likely that businesses will challenge a forced disclosure requirement as they have done with individual state legislation regarding genetically modified labels and release of toxic chemicals. *See* Gary D. Bass, *Big Data and Government Accountability: An Agenda for the Future*,

Although the Authors believe reducing negative rent seeking and data inequality are significant governmental interests, any legislation in this area would need to be narrowly tailored and likely would be limited to advising users of the use and aggregation of their data along with their ability to opt out.

Data brokers' intellectual property concerns with respect to sharing their underlying data could be alleviated through the execution of data use agreements similar to those in federally funded and restricted research relationships.[369] The proposed data use agreements could require a formal review process comprised of both public and private stakeholders to identify the need for the information and why it is unavailable elsewhere. The agreement could also limit the data's reuse without the data broker's prior consent. If data use agreements or disclosure legislation are not feasible, there are additional mechanisms for furthering data equality, such as tax incentives, public information campaigns on the diminished credibility of research that lacks transparency, and public-private partnerships with data brokers like Acxiom. Because Acxiom has acknowledged the inaccuracies inherent in data gathering, it may be willing to explore whether more transparency in research is warranted, highlighting the confidence in its data-gathering techniques and allowing information derived from them to be challenged and openly corroborated.[370] A trend in favor of those data brokers with transparent data through industry best practices should be promoted—such as allowing those that comply to tout their transparency with a certificate and trademark regarding their independence and transparency, like "Transparent Data."[371]

Finally, and most importantly, one avenue for reducing data inequality is for the federal government to assess privacy concerns in a broader context, limiting discretionary FOIA denials.[372] For these and

---

11 I/S: J.L. & POL'Y FOR INFO. SOC'Y 13, 21–23, 37 (2015) (arguing for a proactive disclosure requirement for the government to follow); *see also* Bradford W. Hesse, Richard P. Moser & William T. Riley, *From Big Data to Knowledge in the Social Sciences,* 659 ANNALS AM. ACAD. POL. & SOC. SCI. 16, 19–21 (2016) (explaining that federally funded research often requires full publication of the research data within twelve months of publication).

369. *See, e.g.*, U.S. DEP'T OF HEALTH & HUMAN SERVS., PRACTICES GUIDE: DATA USE AGREEMENT, https://www2.cdc.gov/cdcup/library/hhs_eplc/55%20-%20Data%20Use%20Agreement%20(DUA)/EPLC_DUA_Practices_Guide.pdf [https://perma.cc/T257-VMCC] (last visited Feb. 12, 2018).

370. *See* Simonds, *supra* note 326.

371. *See, e.g., Acxiom Invites Consumers to Visit Aboutthedata.com™ and Calls for More Transparency on Data Privacy Day*, ACXIOM, https://www.acxiom.com/news/acxiom-invites-consumers-visit-aboutthedata-com-calls-transparency-data-privacy-day/ [https://perma.cc/KYP8-K6V2] (last visited Feb. 12, 2018).

372. *But see generally* David E. Pozen, *Deep Secrecy*, 62 STAN. L. REV. 257 (2010) (discussing the difficulty in accessing governmental information, particularly where the existence

other data releases, the government should consider whether the individual whose privacy is at issue has otherwise released the information through other commercial online means. In this regard, researchers may have more avenues to access accurate and transparent data.

VIII. CONCLUSION

Aggressively maintaining privacy in government data while freely allowing private data sources to enjoy immense benefits without commensurate privacy obligations seems incongruous. Inevitably, the imbalanced burden shifting will result in *data inequality*,[373] whereby those with the resources have access to data, and the rest of the public will have little or no access to meaningful data. The Authors argue for a reduction in data inequality and the proper balancing of the government's privacy obligations compared to the data brokers' infinite access to data. Although regulatory reform in this area is unlikely during the Trump administration, the government can ensure it releases more records within its discretionary release capabilities, and it can incentivize data brokers to release necessary data for research. Additionally, it could implement an informational campaign regarding the information the data brokers gather and the lack of credibility that any research has without the ability to cross-check the underlying data provided by data brokers. Ultimately, a combination of solutions similar to the recommendations herein would reduce the growing data inequality and limit any concomitant negative rent-seeking effects. Although society has benefited from the positive aspects of technology, one must query whether "benefit" is being accurately interpreted in light of the data brokers' rent-seeking behavior. It would be more prudent to heed the warning "To Serve Man, it's . . . it's a cookbook!"[374]

---

of the information is hidden, and examining the theories behind secrecy); Samaha, *supra* note 32 (detailing the negative consequences surrounding one's ability to access too much public information).

373. *See* discussion *supra* Part VII.

374. This refers to a short story that became Episode 89 of the Twilight Zone airing on CBS, March 2, 1962, in which aliens provide Earth with advanced technology that ends hunger, among other benefits, and leave behind a book that says the aliens are "to serve man." However, upon visiting the aliens' planet, it becomes evident that "To Serve Man" is actually a cookbook on how to cook humans. *See* DAMON KNIGHT, TO SERVE MAN (1950); *see also The Twilight Zone: To Serve Man* (CBS television broadcast Mar. 2, 1962).