MEASURING AND PROTECTING PRIVACY IN THE ALWAYS-ON ERA

Dan Feldman† *& Eldar Haber*††

ABSTRACT

Data-mining practices have greatly advanced in the interconnected era. What began with the internet now continues through the Internet of Things (IoT)—whereby users can constantly be connected to the internet through various means, like televisions, smartphones, wearables, and computerized personal assistants, among other "things." As many of these devices constantly receive and transmit data, the increased use of IoT devices might lead society into an "always-on" era, where individuals are "datafied"—constantly quantified and tracked.

This situation leads to difficult policy choices. On the one hand, the current sectorial regulatory approach, which protects privacy through regulating information gathering or use only in pre-defined industries or specified cohorts, greatly risks individuals' privacy. On the other hand, strict privacy regulations might diminish data utility, which is crucial for technological development and innovation. There is a tradeoff between data utility and privacy protection, and the sectoral approach to privacy does not strike the right balance. This Article proposes a technological solution that might help. Relying on a method called differential privacy, this Article suggests adding "noise" to data deemed sensitive ex ante. In short, combining computational solutions with formulas that measure the probability of data sensitivity will better protect privacy in the always-on era.

This Article introduces legal and computational methods that could be used by IoT service providers and can optimally balance the tradeoff between data utility and privacy. Part II discusses the protection of privacy under the sectoral approach and estimates what values this approach embeds. Part III discusses privacy protection in the "always-on" era. This Part assesses how technological changes have shaped the sectoral regulation regime, then discusses why IoT devices negatively impact privacy, and finally explores the potential regulatory mechanisms that might meet the challenges of the "always-on" era. After concluding that the current regulatory framework is severely limited in protecting individuals' privacy, this Article discusses technology as a solution in Part IV. This Part proposes a new computational model that relies on differential privacy and a modern invention called private coresets. This proposed model introduces "noise" to users' data according to the probability that the IoT device collects sensitive data, in order to preserve individuals' privacy and ensure service providers can utilize the data at the same time.

DOI: https://doi.org/10.15779/Z38HH6C63R

† Senior Lecturer, Computer Science Department, University of Haifa; Director, Robotics & Big Data Lab, University of Haifa; Faculty member, Center for Cyber, Law and Policy (CCLP), University of Haifa.

†† Senior Lecturer, Faculty of Law, University of Haifa; Visiting Professor, Bocconi University, Italy; Faculty member, Center for Cyber, Law and Policy (CCLP), and Haifa Center for Law and Technology (HCLT), University of Haifa. This work was supported by the Center for Cyber Law & Policy at the University of Haifa in conjunction with the Israel National Cyber Directorate in the Prime Minister's Office.

TABLE OF CONTENTS

I. INTRODUCTION

Technology has posed many threats to individuals' privacy throughout history. Digitization further expanded the risks to privacy by facilitating an increase in data mining and storage capacities. Yet, the internet is not the most threatening technological innovation to privacy, as a newer technological innovation might increase such risks substantially. In what is termed the Internet of Things (IoT)—where ordinary household objects become computerized and connected to the internet—data collection and retention capabilities increase dramatically. This change has enabled service providers to collect massive amounts of sensitive data about their users. IoT devices might capture, to name but a few examples, conversations, imagery, videos, geolocation, biometric data, and even vital signs (e.g., blood pressure or heart rate).[1]

Historically, Congress was quite responsive to technological inventions and digitization that potentially threatened individuals' privacy. Beginning in the 1970s, Congress reacted to privacy threats by crafting a series of federal laws that protect privacy within specified industries or cohorts. These laws can be categorized as protecting financial privacy, educational privacy, health

1. *See* Paul Ohm, *Sensitive Information*, 88 S. CAL. L. REV. 1125, 1143–44 (2015).

privacy, children's privacy, and consumer data privacy.[2] Under this so-called sectoral approach to privacy, regulation applies to a specific context of information gathering or use and is directed at specific pre-defined industries or specific cohorts.[3]

However, the emergence of IoT might make such a sectoral approach to privacy obsolete. IoT is driving society into an "always-on" era in which, more so than ever before, individuals are constantly surrounded by devices that capture their daily routines, including highly sensitive data. Consequently, it is problematic that the current sectoral privacy protection approach does not generally apply to the service providers that offer IoT services; rather, it applies based on the type of data or sector. This leaves IoT users no proper safeguards against such datafication. Thus, IoT exacerbates the limitations of the sectoral approach in the "always-on" era.

Protecting privacy in the always-on era necessitates rethinking the sectoral approach altogether. But before implementing non-sectoral, strict privacy regulatory interventions, policymakers must carefully balance the legitimate interests of IoT companies and users. While marketing is often cited as one of the main reasons for comprehensive data mining, IoT companies rely on data for various other purposes, such as the development of their services. But users' privacy should not be abandoned to accommodate the companies' needs. In other words, decision makers must find a proper way to ensure both data utility and users' privacy.

Technology can be the panacea for a proper tradeoff. While technological solutions, such as de-anonymization or encryption, proved insufficient to protect privacy in the past, other solutions could prove otherwise. This Article proposes a new mathematical model, relying mostly on a method called differential privacy.[4] This new model introduces "noise" that hides information about individual users in data deemed "sensitive" ex ante, depending on various parameters, such as the type of the IoT device, the sensors on the device, the types of data gathered, and the ways data is used. In other words, using technology, this Article proposes a mathematical solution that can both aid in protecting the values embedded in the sectoral approach[5] and ensure extensive privacy protection across IoT devices without sacrificing the utility of the data.

---

2. *See infra* Part II.
3. *Id.*
4. *See infra* Section IV.B.
5. The sectoral approach particularly protects financial privacy, educational privacy, health privacy, children's privacy, and consumer data privacy. *See infra* Section II.B.

Part II discusses the protection of privacy under the sectoral approach and extracts the values embedded in that approach. Part III discusses privacy protection in the always-on era. It assesses how technological changes have shaped sectoral regulation, why privacy is negatively impacted by IoT devices, and whether new regulatory mechanisms to solve the challenges arising in the always-on era are viable. After showing that the current regulatory framework is severely limited in protecting individuals' privacy, Part IV discusses the use of technology as a panacea and presents a new mathematical model that relies mostly on differential privacy. The proposed model introduces "noise" into users' data to preserve individuals' privacy—based on the probability of data sensitivity of the IoT device—while enabling service providers to utilize the data. Part V ends by suggesting that any privacy model, including the proposed model in this Article, must be further examined and recalibrated to embed the values that society wishes to protect.

II. THE SECTORAL PRIVACY PUZZLE

Many scholars have attempted to articulate the need to protect privacy and what it should stand for,[6] but privacy has no clear or single definition.[7] The modern view traces back to Samuel Warren and Louis Brandeis, who defined the right to privacy as the "right to be let alone."[8] In time, privacy literature came to deal extensively with forming a theoretical conception of privacy that furnished a better understanding of that right. Alan Westin offered the "control theory," which conceptualizes privacy as the right to control information about oneself.[9] Ruth Gavison and Anita Allen conceptualized privacy within a "limited access theory," which posits that privacy is "related to our concern over our accessibility to others."[10] Finally, Helen Nissenbaum proposed a conceptual framework of privacy as contextual integrity that links the protection of personal information to the norms of specific contexts.[11]

---

6. *See, e.g.*, William L. Prosser, *Privacy*, 48 CALIF. L. REV. 383, 389 (1960) (offering four types of privacy invasions); Daniel J. Solove, *A Taxonomy of Privacy*, 154 U. PA. L. REV. 477 (2006) (offering a framework for a better understanding of privacy).

7. *See* Daniel J. Solove, *Conceptualizing Privacy*, 90 CALIF. L. REV. 1087, 1090 (2002).

8. Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 193 (1890).

9. *See* ALAN F. WESTIN, PRIVACY AND FREEDOM 7 (1967) ("[T]he claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about themselves is communicated to others.").

10. Ruth Gavison, *Privacy and the Limits of Law*, 89 YALE L.J. 421, 523 (1980). *See generally* ANITA L. ALLEN, UNEASY ACCESS: PRIVACY FOR WOMEN IN A FREE SOCIETY (1988).

11. *See generally* HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE (2010).

Although the scholarly debate on how best to articulate the right to privacy continues today, American policymakers formally acknowledged the existence of this right only after Warren and Brandeis' outcry. Upon calling attention to privacy, policymakers inspired broader interest in recognizing a right to it, and along with the influence of William Prosser, common law tort actions that protect privacy emerged.[12] Subsequently, courts also acknowledged privacy rights in various areas of decision making, usually related to intimately personal matters, such as certain reproductive rights.[13] States have also enacted their own privacy laws.[14] But privacy rights remained unrecognized at the federal level until the 1970s, when Congress began enacting legislations aimed at protecting privacy in particular industries or specific contexts. This regulatory approach has come to be known as the "sectoral approach."

A. PRIVACY AND DATA PROTECTION IN THE UNITED STATES

The American protection of information privacy—also known as data protection—mainly adopts the sectoral approach.[15] Unlike an omnibus approach to privacy, embraced by many jurisdictions such as the European Union,[16] a sectoral approach generally protects information privacy only within a specific context of information-gathering or use and is usually directed only to specific pre-defined industries or specific cohorts.

---

12. *See* Prosser, *supra note* 6, at 386–89; *see also* Solove, *supra* note 7, at 1100; RESTATEMENT (SECOND) OF TORTS § 652 (AM. LAW INST. 1965). The four privacy torts that are generally recognized in the United States are intrusion, public disclosure of private facts, false light, and appropriation. Notably, however, tort law is primarily state-legislated, so the recognition of privacy torts could differ between states. *See* DANIEL J. SOLOVE & PAUL M. SCHWARTZ, INFORMATION PRIVACY LAW 77–231 (3d ed. 2009).

13. *See, e.g.*, Griswold v. Connecticut, 381 U.S. 479, 486 (1965) (acknowledging a right to privacy for married couples' use of contraceptives); Eisenstadt v. Baird, 405 U.S. 438, 443 (1972) (protecting the right of unmarried individuals to possess contraception); Roe v. Wade, 410 U.S. 113 (1973) (acknowledging a right to privacy in a woman's decision to have an abortion under the Due Process Clause of the Fourteenth Amendment); Lawrence v. Texas, 539 U.S. 558 (2003) (invalidating sodomy laws due to sexual privacy).

14. *See* Daniel J. Solove & Paul M. Schwartz, *An Overview of Privacy Law*, *in* PRIVACY LAW FUNDAMENTALS 145–47 (2015).

15. The conventional concept of information privacy refers to protecting a right to control one's personal data. Beyond information rights, privacy rights could also be spatial, regarding individual's physical sphere of control, or decisional, regarding control over personal choices. *See* Joel R. Reidenberg, *Privacy Wrongs in Search of Remedies*, 54 HASTINGS L.J. 877, 878–89 (2003); *see also* Jerry Kang, *Information Privacy in Cyberspace Transactions*, 50 STAN. L. REV. 1193, 1202–05 (1998).

16. An omnibus approach refers to "one overarching law that regulates privacy consistently across all industries." *See* Daniel Solove, *The Growing Problems with the Sectoral Approach to Privacy Law*, TECHPRIVACY (Nov. 13, 2015), https://teachprivacy.com/problems-sectoral-approach-privacy-law [https://perma.cc/X5HV-D3YC].

However, American privacy protection is not entirely sectoral. The right to privacy was interpreted as being embedded in the Bill of Rights, perhaps most evidently in the First, Third, Fourth, and Fifth Amendments.[17] In addition, states sometimes protect privacy within their constitutions or simply legislate state privacy statutes.[18] For example, states have used tort law, or legislated specific privacy or data breach notification statutes, to protect privacy.[19] On both state and federal levels, laws to some extent protect the privacy of data flowing through specific channels of communication, regardless of the data's potential sensitivity.[20] For example, some laws set boundaries on engaging in wiretapping, accessing stored communication, and obtaining data from pen register devices.[21] Some laws also regulate when private entities must keep records for investigatory purposes[22] or facilitate governmental investigations.[23] In other instances, privacy is regulated based on the notion of protecting the public from governmental intrusion with respect

17. The First Amendment protects privacy by permitting the right to speak anonymously and freedom of association. The Third Amendment protects privacy by restricting the government from requiring soldiers to reside in people's houses. The Fourth Amendment prevents the government from conducting "unreasonable searches and seizures," which may implicate individuals' privacy interests. The Fifth Amendment protects individuals against self-incrimination. *See* U.S. CONST. amends. I, III–V; *see also* Daniel J. Solove, *A Brief History of Information Privacy Law*, *in* PROSKAUER ON PRIVACY, 1, 1–5 (2006); Solove & Schwartz, *supra* note 14, at 41.

18. *See* Scott A. Sundstrom, *You've Got Mail! (And the Government Knows It): Applying the Fourth Amendment to Workplace E-Mail Monitoring*, 73 N.Y.U. L. REV. 2064, 2076–77 (1998).

19. *See* Solove & Schwartz, *supra* note 14, at 44; Priscilla M. Regan, *Federal Security Breach Notifications: Politics and Approaches*, 24 BERKELEY TECH. L.J. 1103, 1108–12 (2009).

20. *See* Ohm, *supra* note 1, at 1136 (articulating these laws as "protected channel laws").

21. *See* The Communications Act of 1934, Pub. L. No. 73-416, § 605, 48 Stat. 1064, 1103–04 (1934) (regulating the practice of wiretapping limitedly); Omnibus Crime Control and Safe Streets Act of 1968, Pub. L. No. 90-351, 82 Stat. 197, 211 (1968) (codified as amended at 18 U.S.C. §§ 2510–2522 (2012)) (regulating wiretapping); Electronic Communications Privacy Act of 1986, Pub. L. No. 99-508, 100 Stat. 1848 (1986) (codified as amended at scattered sections of 18 U.S.C.) (revising the Wiretap Act and adding two other acts to deal with technological developments: The Stored Communications Act (SCA), which regulates access to both the content and metadata stored by electronic communications services; and the Pen Register Act, which regulates devices that obtain information about calls).

22. *See, e.g.*, Bank Secrecy Act of 1970, Pub. L. No. 91-508, 84 Stat. 1114 (1970) (mandating federally insured banks and other financial institutions to aggregate financial data and report in order to assist law enforcement agencies in conducting financial investigations); *see also* Solove, *supra* note 17, at 1–29.

23. *See, e.g.*, Communications Assistance for Law Enforcement Act (CALEA) of 1994, Pub. L. No. 103-414, 108 Stat. 4279 (1994) (requiring telecommunication providers to facilitate government interceptions of communications and surveillance under some circumstances).

to what data the state is entitled to collect[24] or how state agencies should handle acquired data.[25]

Some forms of privacy protections, although important, are not part of this Article's general evaluation of privacy. First, Constitutional protection is excluded, as it does not relate to the practices of private actors.[26] Second, the states' protection of privacy rights is also excluded, as it is incoherent and would apply inconsistently depending on the individual policymaker. Finally, public data collection and retention protections—in contrast to data collection and retention by private parties—are excluded because they are less relevant to the discussion on IoT, which is a field currently controlled mainly by private companies.

---

24. *See, e.g.*, Foreign Intelligence Surveillance Act (FISA) of 1978, Pub. L. No. 95-511, 92 Stat. 1783 (1978) (regulating foreign intelligence gathering within the United States); Privacy Protection Act of 1980, Pub. L. No. 96-440, 94 Stat. 1879 (codified at 42 U.S.C. § 2000(a)(a) (2012)) (restricting the government's ability from conducting unlawful searches and seizures of work product of the press and media); Electronic Communications Privacy Act of 1986, Pub. L. No. 99-508, 100 Stat. 1848 (1986) (codified as amended at 18 U.S.C. §§ 2510–2522 (2012)) (regulating electronic storage and surveillance); Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001, Pub. L. No. 107-56, 115 Stat. 272 (2001) (codified as amended at scattered sections of U.S.C. (2012)) (amending various acts such as the ECPA and FISA, loosening requirements for data gathering by law enforcement agencies under some circumstances); *see also* DANIEL J. SOLOVE & PAUL M. SCHWARTZ, PRIVACY LAW FUNDAMENTALS 4 (2017). Individuals' privacy interests in personally identifiable information in the possession of federal agencies received further protection. *See* Privacy Act of 1974, Pub. L. No. 93-579, 88 Stat. 1896 (codified as amended at 5 U.S.C. § 552(a) (2012)). The Federal Trade Commission (FTC) is also tasked with protecting consumers from "unfair or deceptive acts" under § 5 of the Federal Trade Commission Act and it generally regulates commercial collection, use, and release of data under some circumstances. *See* The Federal Trade Commission Act (FTC Act), Pub. L. No. 63-203, 38 Stat. 717, 719 (codified at 15 U.S.C. §§ 45(a), 6505(a) (2012)). For more on the FTC's role in the field of data protection, see generally FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE, RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS 30 (Mar. 2012), http://www.ftc.gov/os/2012/03/120326privacyreport.pdf [https://perma.cc/MHB5-R8TX].

25. *See, e.g.*, Computer Matching and Privacy Protection Act of 1988, Pub. L. No. 100-503, 102 Stat. 2507 (1988) (codified at 5 U.S.C. § 552(a) (2012)); *see also* Driver's Privacy Protection Act of 1994, Pub. L. No. 103-322, 108 Stat. 2099 (codified at 18 U.S.C. §§ 2721–2725 (2012)) (regulating the states' authority to disclose personal driver records by prohibiting, with exceptions, the disclosure or sale of drivers records without obtaining prior consent from the individual); Solove, *supra* note 17, at 1–37. Notably, the law also governs privacy through the lens of government records, granting individuals rights regarding their personal data stored in government records systems. *See, e.g.*, Privacy Act of 1974, Pub. L. No. 93-579, 88 Stat. 1896 (codified at 5 U.S.C. § 552(a) (2012)) (regulating certain kinds of collection and use of records by certain federal agencies, excluding the private sector, state, and local agencies); Solove & Schwartz, *supra* note 14, at 26.

26. *See* Reidenberg, *supra* note 15, at 879.

Instead, this Article focuses on the sectoral approach at its core, which is the federal protection of privacy as applied to the private sector. To better understand the sectoral approach, Section II.B first reviews and analyzes the five categories of sectoral privacy according to federal statutes regulating private parties: financial privacy, educational privacy, health privacy, children's privacy, and consumer data privacy.[27]

The first category is financial privacy, where financial information is granted federal protection under some circumstances.[28] The first federal law that regulated private use and dissemination of information was the Fair Credit Reporting Act of 1970 (FCRA).[29] The FCRA generally regulates the use of credit reports, supervising "the collection, maintenance and dissemination of 'consumer reports' "[30] and require "consumer reporting agencies to maintain procedures to ensure 'maximum possible accuracy.' "[31] It was enacted due to privacy concerns regarding "exclusion, secondary use, and disclosure" of data gathered by credit bureaus.[32] Essentially, the FCRA imposes obligations on consumer reporting agencies and provides individuals with certain rights and control over personal financial records held by credit reporting companies.[33]

In 1978, the Right to Financial Privacy Act (RFPA) further regulated financial data.[34] RFPA sets limits on financial institutions' disclosure of

27. It should be further noted that there are federal acts that apply to the private sector, but focus on conducting an activity by a specific entity, which will be excluded from this analysis. One example is the Employee Polygraph Protection Act of 1988 (EPPA), which regulates the use of polygraphs by private employers. *See* Employee Polygraph Protection Act of 1988, Pub. L. No. 100-347, 102 Stat. 646 (1988) (codified at 29 U.S.C. §§ 2001–2009 (2012)). For more on federal privacy acts, see Solove & Schwartz, *supra* note 14, at 42–44.

28. Beyond the sectoral laws discussed in this Part, other federal laws also relate to financial regulation. *See, e.g.*, Bank Secrecy Act of 1970, Pub. L. No. 91-508, 84 Stat. 1114 (1970).

29. *See* Daniel J. Solove & Chris Jay Hoofnagle, *A Model Regime of Privacy Protection*, 2006 U. ILL. L. REV. 357, 359–60 (2006).

30. *Id.* (citing 15 U.S.C. § 1681 (2012)).

31. *Id.* (citing 15 U.S.C. § 1681(e)(b)). The FCRA was amended by the Fair and Accurate Credit Transaction Act of 2003 (FACTA) with the aim to prevent identity theft and promote accurate credit rating by requiring credit reporting agencies to provide consumers with an annual credit report. *See* Fair and Accurate Credit Transaction Act of 2003, Pub. L. No. 108-159, 117 Stat. 1952, 1968 (2003); *see also* Solove & Schwartz, *supra* note 14, at 42–43.

32. Anne Marie Helm & Daniel Georgatos, *Privacy and Mhealth: How Mobile Health "Apps" Fit into A Privacy Framework Not Limited to HIPAA*, 64 SYRACUSE L. REV. 131, 145 (2014).

33. One example is the right of individuals to request a copy of their credit report. *See* 15 U.S.C. § 1681(b) (2012) (noting more permissible purposes of consumer reports).

34. *See* 12 U.S.C. §§ 3401–3422 (2012); Solove, *supra* note 17, at 1–30.

financial records to a government authority without a warrant or subpoena.[35] Under this act, an unauthorized disclosure by a financial institution or by any government agency obtaining financial records could result in civil penalties.[36] In addition, under the financial privacy category one might also include the Financial Services Modernization Act of 1999, also known as the Gramm-Leach-Bliley Act (GLBA).[37] Although the GLBA does not necessarily relate directly to financial data, it generally regulates financial institutions' processing of personal information,[38] concerning the collection, use, and disclosure of personally identifiable financial information.[39] With some exceptions, the GLBA obliges financial services entities to secure customer records, provide notice and opt-out procedures to consumers before sharing their information with some third parties, and disclose their privacy practices.[40]

The second category is educational privacy, which affords legal protection of student information privacy. A key example is the Family Educational Rights and Privacy Act of 1974 (FERPA).[41] FERPA protects the privacy of school records by regulating access to educational records, students' private records, and other information maintained by educational institutes, such as health records, psychological evaluations, and additional information directly related to students.[42] With some exceptions, students or their parents must consent before an institution may hand over personally identifiable information.[43] FERPA also grants parents and students access to students' files, in order to challenge false or harmful information contained in them.[44]

---

35. *See* 12 U.S.C. §§ 3401–3422; George B. Trubow & Dennis L. Hudson, *The Right to Financial Privacy Act of 1978: New Protection from Federal Intrusion*, 12 J. MARSHALL J. PRAC. & PROC. 487, 497 (1979); *see also* Solove, *supra* note 17, at 1–30.

36. 12 U.S.C. § 3417.

37. *See* The Financial Services Modernization Act (Gramm-Leach-Bliley) Act, Pub. L. No. 106-102, 113 Stat. 1338 (1999) (codified as amended at scattered sections of 12 & 15 U.S.C.).

38. *See* Solove, *supra* note 17, at 1–39.

39. Defined as "nonpublic personal information." *See* 15 U.S.C. §§ 6801–6802 (2012).

40. *See* Theodore Rostow, *What Happens When an Acquaintance Buys Your Data: A New Privacy Harm in the Age of Data Brokers*, 34 YALE J. ON REG. 667, 677 (2017). This rationale arose partly from "privacy concerns regarding consumer financial information." Jolina C. Cuaresma, *The Gramm-Leach-Bliley Act*, 17 BERKELEY TECH. L.J. 497, 497 (2002).

41. *See* Family Educational Rights and Privacy Act of 1974, Pub. L. No. 93-380, 88 Stat. 57 (1974) (codified as amended at 20 U.S.C. § 1232(g) (2012)); 34 C.F.R. § 99 (2018).

42. *See* Solove & Schwartz, *supra* note 14, at 42; *see also* Dalia Topelson et al., *Privacy and Children's Data—An Overview of the Children's Online Privacy Protection Act and the Family Educational Rights and Privacy Act*, 23 BERKMAN CTR. RES. PUB. 1, 2 (Nov. 14, 2013), http://dx.doi.org/10.2139/ssrn.2354339 [https://perma.cc/87PP-UUX7].

43. 34 C.F.R. § 99.31 (2018).

44. *See* Ohm, *supra* note 1, at 1157–58.

Notably, the scope of educational privacy on the federal level is rather limited. FERPA only applies to educational institutions or agencies that receive federal funds from the U.S. Department of Education (DoE).[45] While educational privacy on the federal level is limited in scope, the rationale behind FERPA can readily be understood as contextual. This law seeks to protect the confidentiality of certain records accumulated in educational facilities because these institutions collect information that might be highly sensitive.

The third category is health privacy, where health information deserves stronger data protection than other 'regular' data. The first federal acknowledgment of the importance of health data was embedded in the Freedom of Information Act (FOIA).[46] This law generally mandates disclosure of information upon FOIA requests. But it exempts public access to (1) government records for "personnel and medical files and similar files the disclosure of which would constitute a clearly unwarranted invasion of personal privacy,"[47] and (2) "records or information compiled for law enforcement purposes . . . [that] could reasonably be expected to constitute an unwarranted invasion of personal privacy," potentially including health data.[48]

Apart from exempting health information from FOIA requests, Congress afforded more federal protection to health privacy in 1996 due to a perceived need to digitize health information and preserve the confidentiality of such information. The Health Insurance Portability and Accountability Act (HIPAA) of 1996, which regulates privacy in health records, mandates the Secretary of the Department of Health and Human Services (HHS) to promulgate rules that govern how states must protect the confidentiality of certain health information.[49] The initial rationale was not the protection of privacy per se but the creation of new standards with the goal of "reducing administrative costs."[50]

Today, health information is mainly protected on the federal level under what is collectively termed the HIPAA Privacy Rule.[51] Because of a perceived

45. *See* 20 U.S.C. § 1232(g)(a)(1)(A) (2012); Topelson et al., *supra* note 42, at 3. The entities covered by FERPA include elementary and secondary schools, school districts, colleges and universities, and state educational agencies, along with other institutions that provide educational services. *See* 34 C.F.R. § 99.1(a)(1–2) (2018).

46. *See* Helm & Georgatos, *supra* note 32, at 147; *see also* Freedom of Information Act (FOIA), Pub. L. No. 90-23, 81 Stat. 54 (1966) (codified at 5 U.S.C. § 552(a)(3)(A) (2012)).

47. *See* 5 U.S.C. § 552(b)(6); Helm & Georgatos, *supra* note 32, at 147.

48. 5 U.S.C. § 552(b)(7)(C).

49. *See* Solove, *supra* note 17, 1–38.

50. H.R. REP. NO. 104-497, at 61 (1996); Ohm, *supra* note 1, at 1150.

51. *See* Ohm, *supra* note 1, at 1150–51; *see also* Standards for Privacy of Individually Identifiable Health Information, 67 Fed. Reg. 53, 182 (Aug. 14, 2002) (codified at 45 C.F.R. pts. 160 & 164 (2018) (establishing national standards for protecting certain health

need to strengthen enforcement and expand patient rights, Congress further revised HIPAA in 2009 under the Health Information Technology for Economic and Clinical Health Act (HITECH).[52] The HITECH Act broadened HHS's authority to encompass "business associates" and all businesses that receive information from entities covered by HIPAA.[53] It also added a security breach notification provision[54] and dramatically increased the penalties for HIPAA violations. The HIPAA final rule (or Omnibus Rule) was released in 2013.[55]

HIPAA regulation applies to health data held by covered entities or business associates.[56] Health data is any information, oral or recorded, in any form or medium, that is created or received by various defined entities,[57] which

> [r]elates to the past, present, or future physical or mental health or condition of an individual; the provision of health care to an individual; or the past, present, or future payment for the provision of health care to an individual.[58]

---

information); Final Omnibus HIPAA Rule Preamble, 78 Fed. Reg. 5567 (Jan. 25, 2013) (expanding the definition of business associate and the reach of HIPAA). HIPAA was set to come into force in 2000, but did so only on April 14, 2003. Accordingly, the HIPAA Security Rule was finalized in 2003, but compliance was set for April 21, 2005. *See* Daniel J. Solove, *HIPAA Turns 10: Analyzing the Past, Present, and Future Impact*, 84 J. AM. HEALTH INFO. MGMT. ASS'N 22, 24–25 (2013).

52. Health Information Technology for Economic and Clinical Health Act (HITECH), Pub. L. No. 111-5, 123 Stat. 226 (2009) (codified as amended at scattered sections of 42 U.S.C.). This Act "was passed as a subsection of the American Recovery and Reinvestment Act (ARRA) of 2009." Kimberly L. Rhodes & Brian Kunis, *Walking the Wire in the Wireless World: Legal and Policy Implications of Mobile Computing*, 16 J. TECH. L. & POL'Y 25, 40 (2011); *see also* Solove, *supra* note 17, at 26–28.

53. HITECH also promulgated a data breach notification requirement. *See* HITECH, *supra* note 52; *see also* Solove, *supra* note 17, at 26.

54. *See* 45 C.F.R. §§ 164.400–414 (2018).

55. Modifications to the HIPAA Privacy, Security, Enforcement, and Breach-Notification Rules under the HITECH and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules, 78 Fed. Reg. 5566 (Jan. 25, 2013). For more on the HIPAA rule, see Frank Pasquale & Tara Adams Ragone, *Protecting Health Privacy in an Era of Big Data Processing and Cloud Computing*, 17 STAN. TECH. L. REV. 595, 608–20 (2014).

56. *See* 45 C.F.R. §§ 160.102–103 (2018).

57. These entities include health care provider, health plan, public health authority, employer, life insurer, school or university, or health care clearinghouse. *See* 45 C.F.R. § 160.103.

58. *Id.*

HIPAA "require[s] stronger protections and affirmative consumer consent for certain uses of financial and health data, like marketing."[59]

The HIPAA Privacy Rule governs protected health information (PHI), which includes any "individually identifiable health information" that these entities hold, such as demographic data and information relating to a patient's medical background and care.[60] It requires, inter alia, the anonymization of health data by removal of various types of identifiers.[61] The HIPAA Security Rule provides standards for protecting PHI in electronic form that the covered entity "creates, receives, maintains, or transmits."[62] Under HIPAA, covered entities are required to: (1) designate a privacy official and develop and implement privacy policies; (2) ensure that only the "minimum necessary PHI be accessed and used" and that people authorize disclosure of their PHI (with few exceptions); (3) provide patients with a set of rights; and (4) mandate security safeguards.[63]

The most recent health privacy legislation was passed in 2008: the Genetic Information Nondiscrimination Act (GINA).[64] GINA regulates the use of genetic predisposition to disease by group health plans and health insurers when basing coverage decisions or setting premiums.[65] It also restricts employers from using genetic information when making personnel decisions affecting their employees.[66]

The fourth category is children's privacy, wherein Congress acknowledged the importance of protecting children's privacy online.[67] The Children's Online

---

59. *See* Andrea Reichenbach, *Defining 'Sensitive' in World of Consumer Data*, ACXIOM (July 27, 2015), https://www.acxiom.com/blog/defining-sensitive-world-consumer-data/ [https://perma.cc/RWQ4-SJ33].

60. *See* HIPAA Privacy Rule, 45 C.F.R. § 164.514(b)–(c) (2018); Rostow, *supra* note 40, at 676–77.

61. *See* 45 C.F.R. § 164.514(b)–(c). More closely, the HIPAA Privacy Rule requires that the information will neither identify an individual nor provide "a reasonable basis to believe that the information can be used to identify an individual." § 164.514(a). This could be achieved either by a statistical or a safe harbor standard (which for the latter, must include the suppression or generalization of eighteen enumerated identifiers). *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1737 (2010).

62. 45 C.F.R. § 160.103(4)(i).

63. *See* Solove, *supra* note 17, at 26.

64. Genetic Information Nondiscrimination Act (GINA) of 2008, Pub. L. No. 110-233, 122 Stat. 881 (2008) (codified in scattered sections of 26, 29, and 42 U.S.C.).

65. *Id.* §§ 101–105.

66. *See* Ohm, *supra* note 1, at 1137; GINA, §§ 201–205. For more on GINA, see generally Jennifer J. Lee, Note, *The First Civil Rights Act of the 21st Century: Genetic Information Nondiscrimination Act of 2008*, 4 I/S: J.L. & POL'Y FOR INFO. SOC'Y 779 (2008).

67. It is worth mentioning that Congress also sought to regulate the exposure of children to inappropriate materials online by enacting the Child Online Protection Act, but it eventually

Privacy Protection Act (COPPA) of 1998 regulates the use of children's personal information on the internet.[68] COPPA is supplemented by a rule issued by the FTC known as the COPPA Rule.[69] Both forms of regulation apply to Online Service Providers (OSPs)[70] that target children under age thirteen or knowingly collect personal information from them.[71] They were intended to "prohibit unfair or deceptive acts or practices in connection with

failed to pass constitutional muster because it placed an "impermissible burden" on speech. ACLU v. Reno, 217 F.3d 162, 166, 168–69 (3d Cir. 2000) (referencing The Child Online Protection Act, Pub. L. No. 105-277, 112 Stat. 2681–736 (1998)).

68. *See* Kathryn C. Montgomery & Jeff Chester, *Data Protection for Youth in the Digital Age: Developing a Rights-based Global Framework*, 1 EUROPEAN DATA PROTECTION L. REV. 277, 279–84 (2015). It should be noted that the Family Educational Rights and Privacy Act also regulates children's informational privacy and family privacy. FERPA, however, applies only to the release of educational records to unauthorized persons by educational institutions. *See* The Family Educational Rights and Privacy Act (FERPA), Pub. L. No. 93-380, 88 Stat. 57 (1974) (codified at 20 U.S.C. § 1232(g) (2012)); *Family Educational Rights and Privacy Act (FERPA)* 3 U.S. DEP'T EDUC., http://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html [https://perma.cc/9V2C-JBE2] (last visited Jan. 9, 2020).

69. Children's Online Privacy Protection Rule, 64 Fed. Reg. 59,888 (Nov. 3, 1999) (codified at 16 C.F.R. § 312 (2018)) [hereinafter COPPA Rule]. The COPPA Rule took effect in April 2000 and was last updated in 2013. For the latest update, see 78 Fed. Reg. 3972 (Jan. 17, 2013).

70. COPPA refers to OSPs as "operators" and defines them as

> any person who operates a website [or] online service and who collects or maintains personal information from or about the users of or visitors to such website or online services, or on whose behalf such information is collected or maintained, where such website or online service is operated for commercial purposes, including any person offering products or services for sale through that website or online service, involving commerce.

15 U.S.C. § 6501(2) (2012).

71. "Personal information" is defined as

> individually identifiable information about an individual collected online, including: (1) A first and last name; (2) A home or other physical address including street name and name of a city or town; (3) Online contact information . . . ; (4) A screen or user name where it functions in the same manner as online contact information . . . ; (5) A telephone number; (6) A Social Security number; (7) A persistent identifier that can be used to recognize a user over time and across different Web sites or online services . . . ; (8) A photograph, video, or audio file where such file contains a child's image or voice; (9) Geolocation information sufficient to identify street name and name of a city or town; or (10) Information concerning the child or the parents of that child that [is collected] from the child and combine[d] with [one of the above identifiers].

16 C.F.R. § 312.2 (2018).

personally identifiable information from and about children on the internet," and the FTC enforces both forms of these regulations.[72]

The fifth category is consumer data privacy, which provides that consumer data might be perceived as highly sensitive and hence need firmer protection in some contexts.[73] One example is the Cable Communication Policy Act (CCPA) of 1984.[74] The CCPA requires cable companies to maintain the confidentiality of cable subscribers' records.[75] The law states that cable operators must inform subscribers about the use and assembly of personally

72. 15 U.S.C. §§ 6501–6505; Children's Online Privacy Protection Rule, 64 Fed. Reg. 59,888 (Nov. 3, 1999) (codified at 16 C.F.R. § 312 (2018)); Danielle J. Garber, *COPPA: Protecting Children's Personal Information on the Internet*, 10 J.L. & POL'Y 129, 153 (2002). An "unfair or deceptive" act or practice is a material "representation, omission or practice that is likely to mislead the consumer acting reasonably in the circumstances, to the consumer's detriment" or a practice that "causes or is likely to cause substantial injury to consumers which is not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers or to competition." *See* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 599 (2014). Substantial injury, in this instance, could apply on both financial harms and unwarranted health and safety risks. *Id.*; 15 U.S.C. § 45 (2012) (outlawing unfair methods of competition); Fed. Trade Comm'n v. Info. Search, Inc., Civ. No. 1:06-cv-01099 (D. Md. Mar. 9, 2007) ("The invasion of privacy and security resulting from obtaining and selling confidential customer phone records without the consumers' authorization causes substantial harm to consumers and the public, including, but not limited to, endangering the health and safety of consumers.").

73. It should be noted that other federal laws protect consumers from abusive telecommunication or marketing practices and relate to privacy to some extent. These laws, however, are excluded from this Article's analysis as they do not directly relate to information privacy. *See, e.g.*, Telephone Consumer Protection Act (TCPA) of 1991, Pub. L. No. 102-243, 105 Stat. 2394 (1991) (codified at 47 U.S.C. § 227 (2012)) (regulating the collection and use of telephone numbers); Controlling the Assault of Non-Solicited Pornography and Marketing (CAN-SPAM) Act of 2003, Pub. L. No. 108-187, 117 Stat. 2699 (2003) (codified at 15 U.S.C. § 7701 (2012)) (regulating the collection and use of email addresses and restricts knowingly sending commercial messages to deceive or mislead recipients); The Junk Fax Prevention Act (JFPA) of 2005, Pub. L. No. 109-21, 119 Stat. 359 (2005) (codified at 47 U.S.C. § 609) (expanding the scope of liability for sending junk fax).

74. Cable Communications Policy Act of 1984, Pub. L. No. 98-549, 98 Stat. 2779 (1984) (codified at 47 U.S.C. § 551 (2012)). Notably, other federal and state laws also address consumer data privacy. Some types of customer information, termed "Customer Proprietary Network Information" (CPNI), regulate the use of information that relates to

> the quantity, technical configuration, type, destination, location, and amount of use of a telecommunications service subscribed to by any customer of a telecommunications carrier, and that is made available to the carrier by the customer solely by virtue of the carrier-customer relationship; and information contained in the bills pertaining to telephone exchange service or telephone toll service received by a customer of a carrier.

47 U.S.C. § 222(h)(1) (2012); *see also* Paul M. Schwartz, *Preemption and Privacy*, 118 YALE L.J. 902, 924–25 (2009).

75. *See* Solove, *supra* note 17, at 1–33.

identifiable information to be collected by the company and the ways in which said information may be disclosed. The law also states specific purposes for which a cable operator may disclose or make use of personal information.[76]

Another example is the Video Privacy Protection Act (VPPA) of 1988.[77] Allegedly, the VPPA was enacted in response to the hearing of the Supreme Court nominee Robert Bork, after reporters tried to obtain the nominee's video rental history list.[78] The VPPA regulates the use of video rental information and generally prohibits "video tape service providers"[79] from disclosing personally identifiable information regarding rent or sale of video material of their customers to a third party.[80] Essentially, it limits some entities' disclosure of forms of video viewing habits.[81] Notably, the VPPA was amended in 2013 in light of new rental services, such as Netflix, ultimately weakening its privacy protections.[82]

In sum, the sectoral approach seeks to protect sensitive data within specified contexts. To understand how technological innovations might challenge the effectiveness of this approach, it must be broken down. A taxonomic analysis will shed some light on the core values and interests that Congress sought to protect with this approach and on new potential challenges to it.

B. SECTORAL PRIVACY IN CONTEXT

The rationales behind the sectoral privacy laws suggest that Congress legislates by seeking to identify which data could become sensitive in some industries or in a specific context and whether the risk of privacy harm could increase due to the context. While sensitivity of information is generally difficult to define, academic scholarship has focused mainly on four factors: "possibility of harm; probability of harm; presence of a confidential

76. Federal law also protects the privacy of satellite subscribers. *See* 47 U.S.C. § 338(i) (2012).

77. *See* Video Privacy Protection Act Amendments Act of 2012, Pub. L. No. 112-258, 126 Stat. 2414 (2012) (codified at 18 U.S.C. §§ 2710–2711 (2012)). For further information on the VPPA, see Schwartz, *supra* note 74, at 912.

78. Solove, *supra* note 17, at 1–34; Ohm, *supra* note 1, at 1140.

79. Notably, "video tape service provider" could be broadly interpreted to extend beyond the traditional video tape services, for example, to DVD service providers. Schwartz, *supra* note 74, at 912.

80. Personally identifiable information is defined as "information which identifies a person as having requested or obtained specific video materials or services from a video tape service provider." 18 U.S.C. § 2710(a)(3) (2012).

81. The sensitivity in this instance relates to the "title, description, or subject matter of any video tapes or other audio visual material." § 2710(b)(2)(D)(ii).

82. *See* Video Privacy Protection Act Amendments Act of 2012, Pub. L. No. 112-258, 126 Stat. 2414 (2012); Ohm, *supra* note 1, at 1141.

relationship; and whether the risk reflects majoritarian concerns."[83] Without normatively evaluating whether and how American policymakers should define sensitive information, this Section strives to locate the current rationales behind the sectoral privacy regulations and the core values that Congress sought to protect using the sectoral approach.

As Section II.A above shows, Congress legislated to protect information privacy in five specific contexts. For instance, children's data is afforded protection online because children are considered a cohort entitled to special care and assistance, and because the internet created new risks for them.[84] Similarly, financial transactions, educational records, health data, and some types of consumer data are considered sensitive enough to warrant protection when the probability of data retention by private parties is deemed high due to the context of its gathering.

Although Congress is not precluded from providing future federal information privacy protection in contexts other than the five sectors listed above, the current legal framework is rather limited in scope. Financial privacy, for instance, is only protected on the federal level when it is gathered by predefined private entities, all of which are, to some extent, related to the financial industry.[85] And health privacy is protected in the context of sensitive information usually collected by the protected institutions (e.g., health records and psychological evaluations). Similarly, educational privacy is protected only when there is a high probability of implicating sensitive data. Although children's privacy is often vulnerable to websites' data-mining practices, federal protection does not extend to offline activities, and even online, it is still limited in scope. Finally, consumer data privacy is only deemed sensitive in a specific context in some industries (e.g., cable subscribers' records and the rent or sale of video materials), but not in other instances.

One might argue that, at root, the federal sectoral approach is reactive in nature, since Congress normally intervened to regulate information privacy when a new challenge arose. Especially when a new technology threatened information privacy, or perhaps when a new technology developed at an intolerable rate, Congress would search the context in which individuals'

---

83. *See* Ohm, *supra* note 1, at 1161.

84. For more on online risks to children, see John Palfrey et al., *Enhancing Child Safety and Online Technologies: Final Report of the Internet Safety Technical Task Force*, BERKMAN CTR. FOR INTERNET & SOC'Y (2008), https://cyber.harvard.edu/sites/cyber.law.harvard.edu/files/ISTTF_Final_Report.pdf [https://perma.cc/WDL4-XME4].

85. *See* Solove & Schwartz, *supra* note 14, at 28; ROBERT ELLIS SMITH, BEN FRANKLIN'S WEB SITE: PRIVACY AND CURIOSITY FROM PLYMOUTH ROCK TO THE INTERNET 316–19 (2000); Solove & Hoofnagle, *supra* note 29, at 359.

privacy was at risk and then regulate the affected industries respectively.[86] Ultimately, Congressional legislations granted protection through the lens of the industry—or the cohort—most affected by a new technology at issue.

Naturally, the sensitive data that Congress sought to protect generally could be gathered by various industries simultaneously and was not fully protected by the current regulatory framework. Thus, it is false to assume that the protected information is only sensitive within these sectors, or that the potential harm to data subjects is only plausible within the specified contexts. And the risk of harm is even higher than before in multiple contexts due to rapid technological developments. So, the sensitivity of data must constantly remain on policymakers' agenda. IoT, perhaps the most dramatic technological innovation in recent years, especially necessitates overall scrutiny of data protection in the United States and a search for other viable solutions.

As the next Part shows, the types of information that should be deemed sensitive and the types of information privacy that deserves protection can swiftly change due to constant innovations in how data are gathered, processed, and stored. But before embarking on a normative evaluation of the new threats to information privacy posed by new technologies, it is essential to discern the role of technology in the development of sectoral privacy protection to date.

III. PROTECTING PRIVACY IN AN ALWAYS-ON ERA

The evolution of technology in the twenty-first century is likely the most rapid in human history. But even prior to these new rapid developments, evolving technology at the time fulfilled a substantial role in establishing sectoral privacy protection. Today, the federal sectoral approach seems ever reactive to new technologies, and it has not responded to the rapid pace of technological development in recent years, including IoT.

To better understand the potential risks that IoT might impose on the right to information privacy and to see if legal intervention could substantially reduce such risks, this Part will proceed as follows. Section A briefly discusses technology's impact on shaping sectoral privacy prior to IoT. Section B introduces the so-called "always-on" era, and the challenges this era poses in the context of sectoral privacy. Section C examines whether, and to what extent, legal intervention might help protect individuals' privacy in this era.

86. *See supra* Section II.A.

A. SECTORAL PRIVACY AND TECHNOLOGY

Technology and privacy protection go hand in hand in America. For example, since people began to communicate via mail and telegraph, concerns over insecurity, disclosure, and breach of confidentiality led many U.S. policymakers to strengthen mail security.[87] Another example is Warren and Brandeis's influential article "The Right to Privacy," which was allegedly inspired by the combination of relatively new technologies, such as the yellow press and cameras (instantaneous photography),[88] along with new business models of some industries that found profitability in publishing such data.[89] Likewise, the spread of telephone use raised various privacy concerns that eventually led to diverse legislative responses on both the state and the federal levels.[90]

Compared to other technological inventions to date, digitization has probably influenced privacy protection the most. In its early days, digitization created the need to develop standards of fair information practices in dealing with citizens' personal information.[91] When electronic communications presented privacy protection with new challenges,[92] Congress passed legislation relating to the interception and access of electronic communications

---

87. *See* Helm & Georgatos, *supra* note 32, at 141–42 (describing how a then-emerging technology (the mail) influenced privacy protection).

88. *See* Warren & Brandeis, *supra* note 8; Solove, *supra* note 17, at 1–11; Andreas Busch, *Privacy, Technology, and Regulation: Why One Size Is Unlikely to Fit All*, *in* SOCIAL DIMENSIONS OF PRIVACY: INTERDISCIPLINARY PERSPECTIVES 303, 304–05 (Beate Roessler & Dorota Mokrosinska eds., 2015).

89. Revealing photographs and gossip about individuals' personal lives was profitable mainly for the penny press. *See* Danielle Keats Citron, *Mainstreaming Privacy Torts*, 98 CALIF. L. REV. 1805, 1807 (2010); Warren & Brandeis, *supra* note 8, at 196.

90. *See, e.g.*, The Communications Act of 1934, Pub. L. No. 73-416, 48 Stat. 1064 (1934); Omnibus Crime Control and Safe Streets Act of 1968, Pub. L. No. 90-351, 82 Stat. 211 (1968) (codified as amended at 18 U.S.C. §§ 2510–22 (2012)); Helm & Georgatos, *supra* note 32, at 142–43.

91. The first American acknowledgment of fair information practices standards was by the Department of Health and Human Services, which in 1973 elaborated a code of practice for the fair treatment of citizens' personal information. *See* U.S. DEP'T OF HEALTH, EDUC. & WELFARE, SECRETARY'S ADVISORY COMM. ON AUTOMATED PERSONAL DATA SYSTEMS, RECORDS, COMPUTERS AND THE RIGHTS OF CITIZENS (1973), *reprinted in* U.S. PRIVACY PROTECTIONS STUDY COMMISSION, PERSONAL PRIVACY IN AN INFORMATION SOCIETY 15 n.7 (1977); Reidenberg, *supra* note 15, at 879–80. For criticism on fair information practices in the United States, see Omer Tene, *Privacy Law's Midlife Crisis: A Critical Assessment of the Second Wave of Global Privacy Laws*, 74 OHIO ST. L.J. 1217, 1218–20 (2013); Joel R. Reidenberg, *Setting Standards for Fair Information Practice in the U.S. Private Sector*, 80 IOWA L. REV. 497, 499–500 (1995).

92. *See* Helm & Georgatos, *supra* note 32, at 143.

and computer tampering.[93] HIPAA, as a final example, was formed under a perceived need to digitize health information and better protect such information due to its potential sensitivity.[94]

As Section II.A implies, digital networks—especially the internet—led to the passage of several Acts designed to better protect individuals' privacy in specific contexts. For instance, COPPA protection for children's privacy online arose out of the potential risks to children's privacy when children surf the web.[95] Arguably, children's information needed as much protection before the internet era as it does now.[96] But before the internet, there were physical barriers to collecting information about children. Therefore, although the need to protect such information existed then, it was simply not on the agenda, either because it was somewhat tricky to violate children's privacy, or perhaps because any potential violation was at a socially tolerable level as conceived by policymakers. However, with the rise of the internet, the ease of conveying information to websites, especially those directed at children, has made the protection of children's information more relevant and crucial—so, COPPA was born.

Regarding the importance of digitization for information privacy, the internet has clearly influenced the development of sectoral privacy laws. But examining the vast amount of sensitive information extracted online,[97] one might conclude that Congress has done little to regulate it in this regard. Considering OSPs' capacity to harvest data online, sectoral privacy clearly has hardly imposed any obligations on OSPs online as it did in regulated sectors of the kinetic world. For instance, if the VPPA and the CCPA were crafted to protect consumers from revealing their preferences or habits regarding what they acquire because such information is sensitive, then online services like

93. *See* Electronic Communications Privacy Act of 1986, Pub. L. No. 100-618, 102 Stat. 3195 (1986) (codified as amended at 18 U.S.C. § 2710 (2012)); The Computer Fraud and Abuse Act, Pub. L. No. 99-474, 100 Stat. 1213–16 (1986) (codified at 18 U.S.C. § 1030 (2012)).

94. *See supra* Section II.A. Other instances are the collection and retention of sensitive consumer data like cable subscribers' records and video rental information as protected by the CCPA and VPPA, respectively. *Id.*

95. For further reading on the rationales behind COPPA, see generally Eldar Haber, *Toying with Privacy: Regulating the Internet of Toys*, 80 OHIO ST. L.J. 399 (2019).

96. It should, however, be further noted that various factors could have also affected the necessity to protect children's rights, and thus, the importance of protecting children might have changed throughout time. One example would be the international acknowledgment of granting such protection in 1989. *See* G.A. Res 44/25, Convention on the Rights of the Child (Nov. 20, 1989). For more on online risks to children, see Palfrey et al., *supra* note 84.

97. Almost everything end-users do on computerized networks is known to private parties. *See*, *e.g.*, Daniel J. Solove, *Digital Dossiers and the Dissipation of Fourth Amendment Privacy*, 75 S. CAL. L. REV. 1083, 1084 (2002).

YouTube and Netflix must comply as well.[98] Moreover, scrutiny of internet usage, such as search query information, could reveal a great deal about individuals and thus represents a major threat to privacy under the sectoral approach, which does not generally apply to OSPs' harvesting personal data online.[99]

But the internet is simply the beginning in this context. Even the ability to collect sensitive information over the internet, no matter how massive it seems, might still be less powerful than that of impending technological innovations. As we move toward an era where ordinary devices, or "things," are becoming interconnected through the internet and are equipped with powerful sensors capable of capturing conversations, imagery, videos, geolocation, biometric data, and even vital signs, such as blood pressure or heart rate, information privacy is at great risk.[100]

Therefore, IoT must be further scrutinized for a grasp of its potential ramifications regarding information privacy. To gain a better understanding of these risks and potential solutions, the next Section will introduce the "always-on" era and analyze sectoral privacy within that context.

B. PRIVACY IN THE "ALWAYS-ON" ERA

Data collection and retention have been constantly on the rise since the invention of the internet.[101] It has enabled both private and public parties to collect massive amounts of data about their users.[102] The internet began to expand beyond traditional computers when other electronic devices, such as phones, TVs, watches, and even homes, suddenly became "smart." Using these smart devices quickly became the norm for many individuals in today's digital society.[103] Not long thereafter, other physical items—or simply

98. *See* Patricia L. Bellia, *Federalization in Information Privacy Law*, 118 YALE L.J. 868, 874–85 (2009).

99. *See* Paul Ohm, *The Rise and Fall of Invasive ISP Surveillance*, 2009 U. ILL. L. REV. 1417, 1417 (2009) (discussing the threats to privacy that arise from ordinary internet usage).

100. *See* Ohm, *supra* note 1, at 1143–44. It should be noted that geolocation information is generally not protected under the sectoral approach, rather, only under COPPA or regarding governmental access to information. *See* 16 C.F.R. § 312.2 (2018).

101. For more on the history of the public internet, see generally Jonathan Zittrain, *A History of Online Gatekeeping*, 19 HARV. J.L. & TECH. 253 (2006).

102. *See* Ben Popken, *Google Sells the Future, Powered by your Personal Data*, NBC NEWS (May 10, 2018), https://www.nbcnews.com/tech/tech-news/google-sells-future-powered-your-personal-data-n870501 [https://perma.cc/2SMD-NUVM].

103. *See, e.g.*, Larry Downes, *Why you may have good reason to worry about all those smart devices*, WASH. POST (Dec. 6, 2016), https://www.washingtonpost.com/news/innovations/wp/2016/12/06/why-you-may-have-good-reason-to-worry-about-all-those-smart-devices/?noredirect=on&utm_term=.f6d8fcb1e7c5 [https://perma.cc/8N7Z-UWG6].

"things"—also emerged as interconnected. Not surprisingly, this technology is hence termed the Internet of Things, or IoT.

IoT undoubtedly increases the possibility of data gathering, in both the types of data and their potential volume, and it could signal a step up in a new generation of data mining.[104] These "things" are capable of gathering massive amounts of data about their users; for example, smart TVs, refrigerators, and even smart washing machines can collect, analyze, and retain data on their users' habits.[105] Smart TVs can also listen to, record, and send to a third party whatever their microphones catch[106] and can even acquire data from a built-in camera.[107] Smartphones in particular enable information gathering of various types by various service providers.[108]

In the development of IoT, an emerging generation of technology could further elevate data collection: devices that operate in an always-on mode, meaning they can constantly collect data even without being active.[109] The definition of always-on devices is largely self-explanatory: such devices either

---

104. *See* Rostow, *supra* note 40, at 686.

105. *See* Chris Hoffman, *How to Stop Your Smart TV From Spying on You*, HOW-TO GEEK (Nov. 16, 2015), https://www.howtogeek.com/233742/how-to-stop-your-smart-tv-from-spying-on-you [https://perma.cc/KD2N-5LWA]. Some Smart TVs could also transmit the names of files on USB drives connected to the television and capture data from networks to which they are attached. *See* Joseph Steinberg, *These Devices May Be Spying on You (Even in Your Own Home)*, FORBES (Jan. 27, 2014), http://www.forbes.com/sites/josephsteinberg/2014/01/27/these-devices-may-be-spying-on-you-even-in-your-own-home/#15407ce56376 [https://perma.cc/Q5WK-9RN3].

106. *See* April Glaser, *Philip K. Dick Warned Us About the Internet of Things in 1969*, SLATE (Feb. 10, 2015), http://www.slate.com/blogs/future_tense/2015/02/10/philip_k_dick_s_1969_novel_ubik_on_the_internet_of_things.html [https://perma.cc/A6DR-FE98].

107. *See* Steinberg, *supra* note 105.

108. For instance, cellular providers could track information about their users, such as with whom the user communicates and where the user goes; manufacturers and providers of software for smartphones, such as Google (Android phones) and Apple (iPhones), could track the actions their users are taking on their phone; and app developers often use their installed apps to extract information from the phone's contact list, microphone, and camera. *See id.* In fact, many flashlight apps gained a reputation of data exfiltration. *See* Robert McMillan, *The Hidden Privacy Threat of...Flashlight Apps?*, WIRED (Oct. 20, 2014), https://www.wired.com/2014/10/iphone-apps [https://perma.cc/J8PS-AC94].

109. There is, however, a difference between always-ready and always-on statuses. Always-ready devices usually process locally to detect a "wake phrase," which triggers the device to begin transmitting data. But always-on devices transmit data all the time while the processing occurs only externally. For the purposes of this Article, always-ready devices count as always-on, since always-ready devices are constantly awaiting the trigger phrase, they must always be "on" and thus could potentially transfer and collect data constantly. For more on this categorization, see *Microphones & the Internet of Things*, FUTURE PRIVACY F. (Aug. 2017), https://fpf.org/wp-content/uploads/2017/08/Microphones-Infographic-Final.pdf [https://perma.cc/EUM2-D6QG].

always await a trigger phrase to begin operating at any moment ("always-ready") or operate constantly without a moment of idleness ("always-on").[110]

Examples of always-on devices can be found in many areas, from young kids using smart, connected toys and devices to individuals (young and old) using computerized personal assistants or operating a smart home.[111] Their functionality and operation can be exemplified through computerized personal assistants like Amazon Echo.[112] For example, Amazon Echo is an "always-ready" device, meaning that it only becomes active upon a voice command like "Alexa" or "Amazon," depending on users' preferences. But for the device to know when the user has operated the activation command, it must constantly await commands by "listening" to its users. Therefore, the device is labeled as always on, even if it presumably deactivates without the command.[113]

These innovative technologies mark the beginning of the always-on era. Due to the devices' mode of operation and their data collection abilities, the

110. For many IoT devices, users simply need to say the voice command to activate them, which Stacey Gray suggested terming as microphone-enabled devices. *See* Stacey Gray, *Always On: Privacy Implications of Microphone-Enabled Devices*, FUTURE PRIVACY F. 3 (Apr. 2016), https://fpf.org/wp-content/uploads/2016/04/FPF_Always_On_WP.pdf [https://perma.cc/QS4Q-KJ8C].

111. Computerized personal assistants are software agents that can perform tasks or services for an individual, usually based on user input, location awareness, and the ability to access information from a variety of online sources. There are various types of computerized personal assistants (e.g., Apple's Siri and Microsoft's Cortana). In 2014, Google even embedded such technology under a pre-installed ability in Google's Chrome browser, which passively listened for the phrase "OK, Google" to launch a voice-activated search function. *See* Tony Bradley, *'OK Google' Feature Removed from Chrome Browser*, FORBES (Oct. 17, 2015), http://www.forbes.com/sites/tonybradley/2015/10/17/ok-googlefeature-removed-from-chrome-browser/#16d299a44e27 [https://perma.cc/8SL7-XFFM]; *see also Top 22 Intelligent Personal Assistants or Automated Personal Assistants*, PREDICTIVE ANALYTICS TODAY, https://www.predictiveanalyticstoday.com/top-intelligent-personal-assistants-automated-personal-assistants/ [https://perma.cc/K2W4-RGPQ] (last visited Jan. 9, 2020).

112. Amazon Echo is "a hands-free speaker you control with your voice." *Amazon Echo*, AMAZON, https://www.amazon.com/Amazon-Echo-Bluetooth-Speaker-with-WiFi-Alexa/dp/B00X4WHP5E [https://perma.cc/9LUQ-FRT3] (last visited Jan. 9, 2020). It "connects to the Alexa Voice Service to play music, make calls, send and receive messages, provide information, news, sports scores, weather, and more—instantly. . . . When you want to use Echo, just say the wake word 'Alexa' and Echo responds instantly." *Id.*

113. Notably, it is difficult to estimate if these devices constantly collect data. Amazon, for instance, claimed that their Echo device only starts recording upon the trigger phrase. Google argues that Google Home (as another example of a computerized personal assistant) only "listens in short (a few seconds) snippets for the hotword. Those snippets are deleted if the hotword is not detected, and none of that information leaves your device until the hotword is heard." *See* Scott Carey, *Does Amazon Alexa or Google Home Listen to My Conversations?*, TECHWORLD (May 25, 2018), https://www.techworld.com/security/does-amazon-alexa-listen-to-my-conversations-3661967 [https://perma.cc/Y4W9-HPHX].

always-on era could lead to the collection and storage of massive quantities of user data of various types. Almost anything could be transmitted and stored,[114] depending mostly on the plausibility of obtaining authorized or unauthorized access to data. It might lead to a constant—and almost endless—collection and retention of data, even when the device seems to have been deactivated.

In the context of sectoral privacy, always-on devices may very well collect and retain information deemed sensitive by Congress and hence should become a subject for sectoral protection under federal laws. Operators of always-on devices might easily cover all the categories of sectoral privacy.[115] Consider, for example, Amazon Echo. If one is present in your household, it can capture any conversation in its vicinity and thus collect data regarding your finance, health, school performance, and any other information that you might consider sensitive. If children are present, it might capture their voices, conversations, questions, and even their musical preferences if they ask the device to play songs. Also, consider an always-on smart TV, or another IoT device equipped with a camera. Besides potentially acquiring watching habits, it could be used or misused to collect sound and imagery from its surroundings—Echo Show and Echo Look are just two examples of devices with a microphone and a camera.[116] Thus, many smart devices can gather almost any information that is already deemed sensitive by Congress. Worse yet, collecting some sensitive data is not just possible; it is highly probable.

To date, this so-called always-on era has had little influence on reforming sectoral privacy, nor modifying it even slightly.[117] When it comes to IoT, sectoral privacy borders on irrelevance. Other than COPPA, which marginally applies to some IoT devices—namely connected smart toys, such as Hello Barbie and My Friend Cayla[118]—most sectoral privacy laws do not apply to

---

114. *See, e.g.*, Nick Ismail, *Storage Predictions: Will the Explosion of Data in 2017 be Repeated in 2018?*, INFORMATION-AGE (Dec. 6, 2017), http://www.information-age.com/explosion-data-2017-repeated-2018-123469890 [https://perma.cc/9YGX-G4YZ].

115. As previously mentioned, sectoral privacy mainly protects financial privacy, educational privacy, health privacy, children's privacy, and consumer data privacy. *See supra* Section II.A.

116. *See Echo Look*, AMAZON, https://www.amazon.com/Amazon-Echo-Look-Camera-Style-Assistant/dp/B0186JAEWK [https://perma.cc/6ESZ-AY7H] (last visited Jan. 9, 2020); *Echo Show*, AMAZON, https://www.amazon.com/Amazon-Echo-Show-Alexa-Enabled-Black/dp/B01J24C0TI/ref=sr_1_1?s=amazon-devices&ie=UTF8&qid=1528381083&sr=1-1&keywords=echo+show&dpID=51syqGPcCmL&preST=_SY300_QL70_&dpSrc=srch [https://perma.cc/49UY-DJAR] (last visited Jan. 9, 2020).

117. While IoT has been debated in Congress, no significant legislation has passed thus far to protect information privacy.

118. Hello Barbie and My Friend Cayla are examples of connected smart toys, or IoToys, which are connected to the internet and can communicate with their users through voice

IoT because they have been crafted very narrowly to address a specific problem. Financial privacy, for instance, will be protected only if the IoT operator is an institution engaging in financial activities, or a certain entity that receives non-public personal information from non-affiliated financial institutions.[119] FCRA will probably not apply unless IoT operators are treated as consumer-reporting agencies.[120] Educational privacy will be generally excluded, as it applies to educational institutes or agencies that receive federal funds from the DoE.[121] Health privacy and consumer data privacy will likewise not be easily protected by these Acts when it comes to IoT devices, as they will not be considered covered entities under federal regulations.[122] All in all, sectoral privacy will not greatly concern IoT.

At root, policymakers must realize that many IoT devices are likely to acquire sensitive data, so privacy protection of relevant data should not be restricted merely to covered entities. Even the context of specific industries (e.g., educational facilities) or a specific population (e.g., young children) is inadequate to protect sensitive information today. When almost everything around us becomes a computer that is connected to the internet,[123] and data become a substantive part of the business model for many companies, the notion of how better to protect users' privacy has to be reconsidered. With few exceptions, most of the core values protected by the sectoral approach could become meaningless with the advance of new technologies, especially IoT. Therefore, protecting sectoral privacy in the always-on era calls for some form of intervention, legal or technological.

C. ALWAYS-ON REGULATIONS

That privacy had met its demise became a popular opinion toward the end of the twentieth century.[124] Some regulators, however, like those of the

commands. For more on the regulation of IoToys in the United States, see generally Haber, *supra* note 95.

119. *See* 12 U.S.C. §§ 3401–3422.

120. *See* 15 U.S.C. § 1681(b) (2012).

121. *See* 20 U.S.C. § 1232(g).

122. *See* 45 C.F.R. §§ 160.102–103.

123. *See generally* BRUCE SCHNEIER, CLICK HERE TO KILL EVERYBODY: SECURITY AND SURVIVAL IN A HYPER-CONNECTED WORLD 5–12 (2018) (describing how IoT turns almost any item into a computer).

124. Many argued that privacy is dead or that it deserves at most minimal protection in the digital age. Others argued that privacy should be treated as a tradeable currency. Scott McNealy, chief executive officer of Sun Microsystems, is quoted as saying, "You have zero privacy anyway . . . . Get over it." Polly Sprenger, *Sun on Privacy: 'Get Over It'*, WIRED (Jan. 26, 1999), https://www.wired.com/1999/01/sun-on-privacy-get-over-it [https://perma.cc/FLA2-7EW5]; *see also* A. Michael Froomkin, *The Death of Privacy?*, 52 STAN. L. REV. 1461, 1462 (2000). For more on the currency argument, see James P. Nehf, *Shopping for Privacy Online:*

European Union, recently made a clear statement regarding privacy under its General Data Protection Regulation (GDPR): protecting privacy still matters, perhaps even more today than ever before.[125] Therefore, private companies should grant their users more effective means of control and protection. While American policymakers still do not protect information privacy as robustly as the European Union does, and perhaps granting such protection might not be achieved easily, they evidently do not disregard this right and instead still seek proper ways to better protect it under their regulatory approach.[126]

Notably, privacy protection does not necessarily require legal intervention. As Lawrence Lessig famously argued, other potential modalities, like the market, social norms, and architecture, could also regulate behavior, with or without the law.[127] The problem with some of these potential modalities in the privacy-protection field and the always-on era lies in their failure to optimally regulate privacy protection on their own. For instance, as history shows, the market as a modality—while arguably an important component of any solution—might be insufficient to regulate privacy due to existing market failures.[128] Likewise, social norms will not effortlessly change the data-mining practices of commercial entities.[129] While this Article considers the potential of both the market and social norms to regulate privacy, it focuses mainly on the modalities of law and technology, probing mainly the law in this Section.

The use of the law as a modality to better protect privacy, by embedding the values protected by the sectoral approach, can take many forms. One might

---

Consumer Decision Making Strategies and the Emerging Market for Information Privacy, 2005 U. ILL. J.L. TECH. & POL'Y 1, 14–17 (2005).

125. *See generally* Commission Regulation 2016/679, 2016 O.J. (L 119) (EU) 1 (repealing Directive 95/46/EC) (General Data Protection Regulation).

126. *See, e.g.*, FED. TRADE COMM'N, INTERNET OF THINGS: PRIVACY AND SECURITY IN A CONNECTED WORLD (Jan. 2015), https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf [https://perma.cc/62U9-6Y6J].

127. *See* LAWRENCE LESSIG, CODE: VERSION 2.0 120–37 (2006); LAWRENCE LESSIG, FREE CULTURE 116–73 (2004) (suggesting four modalities that regulate behavior).

128. To exemplify, many OSPs rely on data as a business model; this can be a market failure if they are a monopoly or operate in an oligopolistic market and thus lack incentives to provide proper privacy protections. Users will generally lack the opportunity to indicate their discontent with such practices. For more on privacy and market failures, see Victoria L. Schwartz, *Corporate Privacy Failures Start at the Top*, 56 B.C. L. REV. 1693 (2016).

129. To name a few examples, social norms generally fail to solve this conundrum, as consumers are generally unaware of data-mining practices; some view privacy as a currency; many fail to understand the implications of data storage; and even those who use these services might not be able to use them in the IoT context. For a general discussion on privacy and social norms, *see* Randall P. Bezanson, *Privacy, Personality, and Social Norms*, 41 CASE W. RES. L. REV. 681 (1991).

argue that to protect privacy properly in the digital age, Congress should choose a different framework (i.e., abandon the sectoral approach entirely), as this does little to advance the rationales behind the current regulatory approach to privacy protection.

Many scholars have long warned how poorly suited the sectoral approach is to protecting privacy in this era.[130] Indeed, it is difficult to grasp why personal information should be treated differently simply because of the identity of the private party that holds it, as the sectoral approach suggests. But it did seem to make sense that your doctor, rather than most individuals you encounter, should obtain your personal information or have full access to your entire medical history. The present nature of data mining could challenge these assumptions. Google most likely has far more intimate and personal information about you than your doctor. It might even know more about you than anyone else in the world, including your family and perhaps even yourself. That is why the sectoral privacy approach might be inadequate and ill-suited to protecting privacy in today's always-on era.

These changes might eventually lead to a comprehensive federal privacy law that could provide a one-size-fits-all approach or create a federal baseline for all industries when dealing with sensitive information.[131] While this remains to be seen, what should be evident in the always-on era is that regulating privacy in industries does little to achieve the goals of the sectoral approach. When the devices that surround us can collect protected forms of information very similar to those of the regulated industries, any sectoral regulation must also apply to OSPs of IoT. The current patchwork regime to protect privacy is thus too outdated to deal with current challenges. Therefore, policymakers are duty-bound to reevaluate the collection, storage, and transfer of information across the private sector, and to regulate it accordingly.

Developing a one-size-fits-all approach does not necessarily mean abandoning the sectoral approach entirely, as other forms of offline data collection might still exist. An extreme ex ante approach, for instance, might argue that the solution for protecting information privacy would be simply to ban data collection and retention in general, at least for some companies or sectors. This general approach, which focuses on how to prevent some forms of data from being retained from the outset, is unsuitable. This is because data serve as a business model for many companies. It is a multibillion-dollar

130. *See, e.g.*, Ohm, *supra* note 61, at 1762; *cf.* Schwartz, *supra* note 74, at 922–31 (discussing the drawbacks of embracing an omnibus privacy regime in the United States).

131. *See, e.g.*, Bellia, *supra* note 98, at 890–900 (advocating for the importance of federalization of information privacy law).

industry with many benefits for its users, such as the offer of free services.[132] Data could be highly valuable for companies, and for some users, as data processing could enable, inter alia, targeted—perhaps more accurate—advertising and suggest personalized services.[133]

Furthermore, the practice of data collection and retention serves many functions and values. It advances knowledge and innovation and is crucial for the development of machine learning, deep learning, and big data analysis, to name but a few examples, all of which rely heavily on large quantities of training data.[134] In addition, various parties, such as credit card companies, use data to reduce exposure to risks and costs of doing business, while they increase companies' effectiveness at raising revenues.[135] In sum, a general approach banning information or simply ignoring the opportunities in data mining altogether is neither practical nor desirable.[136]

But gray areas in such an ex ante approach exist. Depending on various factors regarding information privacy, companies might be allowed to process some forms of data, but not others. They could also be obliged or incentivized to incorporate privacy-enhancing principles into their practices, which could include, inter alia, limits on data collection and retention, data disposal, data accuracy, and various cybersecurity measures—at least for protection against the unauthorized use of these data.[137] In addition, technological developments could aid companies in understanding which data should be retained and which should not. Imagine that your smart device could actually deduce the speaker's identity, and therefore could change privacy settings in keeping with its preferences, or even more closely, protect the privacy of those that Congress sought to protect. So if your smart assistant could differentiate you from your under-thirteen-year-old child, it could potentially retain information

132. *See* Rostow, *supra* note 40, at 687; Cuaresma, *supra* note 40, at 506.

133. Notably, data could be highly valuable for non-profit companies as well, as they might, inter alia, use free open-sourced technology variants of for-profit company technologies in order to service users' device. Some users, however, might view targeted advertising and personalized services as a nuisance.

134. This is especially evident in the context of deep learning, which yields state-of-the-art results in fields such as speech, image, and text recognition, but based on billions of records that were collected from private users. For more on deep learning, see generally Liangpei Zhang et al., *Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art*, 4 IEEE GEOSCIENCE & REMOTE SENSING MAG. 22 (2016); Xue-Wen Chen & Xiaotong Lin, *Big Data Deep Learning: Challenges and Perspectives*, 2 IEEE ACCESS 514 (2014).

135. *See* Cuaresma, *supra* note 40, at 506.

136. *See* Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1, 4–5 (2011).

137. For similar recommendations in the United States, see, for example, *Protecting Consumer Privacy in an Era of Rapid Change*, *supra* note 24, at 15–71.

from you alone and exclude data mining from the child. Developments in voice or facial recognition could advance this rationale.[138]

The problem, however, is that other than children's privacy, protecting sensitive data goes far beyond the speaker's identity. It requires context and evaluation of the data to ascertain whether it is sensitive. It would be very difficult to achieve such a goal ex ante—requiring measures that could accurately predict which information it should not store beforehand.

Even with the invention of such technological measures, or if at least we accept ex ante privacy protection depending on the user's identity, these measures might raise further privacy issues. Embedding technologies like facial or voice recognition in IoT devices might exert a significant negative effect on the right to privacy, as these means rely on biometric features. Even if these features were stored only internally on the device, and even assuming that the biometric data were encrypted against the potential abuse of them, these measures would essentially rely on a form of anonymization, which could easily be de-anonymized, and in consequence users could be re-identified and suffer further damage.[139] For instance, if an Amazon Echo "knows" how to locate persons in a household and retains their preferences, revelation of the household's preferences could easily identify what data were linked to each person. In other words, using this method of protecting one cohort, such as children, might eventually lead to diminished protections for others.

A less extreme ex ante approach could rely on users' preferences (i.e., depending on whether they consent to such potential threats to their privacy in the always-on era). This so-called notice-and-consent mechanism already exists in the United States as part of the Fair Information Practice Principles (FIPPs), which are generally designed to empower data subjects by ensuring that they have sufficient knowledge of a data collector's activities in order to choose to consent to them or not.[140] However, this regulation-by-information approach, which relies on the concept of "informed consent," has proven

138. Facial recognition is also developing at a rapid pace. Both Google Home and Amazon Echo have gained the ability to recognize individual voices by creating a voice profile. *See* Chris Welch, *Amazon's Alexa Can Now Recognize Different Voices and Give Personalized Responses*, VERGE (Oct. 11, 2017), https://www.theverge.com/circuitbreaker/2017/10/11/16460120/amazon-echo-multi-user-voice-new-feature [https://perma.cc/SRN9-JFBX].

139. Paul Ohm categorizes this phenomenon as the "accretion problem," where "[o]nce an adversary has linked two anonymized databases together, he can add the newly linked data to his collection of outside information and use it to help unlock other anonymized databases." Ohm, *supra* note 61, at 1746–48.

140. FIPPs could include, inter alia, notice, choice, access, accuracy, data minimization, security, and accountability. *See* Woodrow Hartzog, *The Inadequate, Invaluable Fair Information Practices*, 76 MD. L. REV. 952, 973–75 (2017).

ineffective in protecting privacy,[141] and likewise from a legal standpoint, individuals are unlikely to challenge potential violation of their privacy for a variety of reasons.[142]

Having determined that an ex ante legal approach is ineffective and insufficient to protect privacy, let us move on to examine ex post approaches. If we allow private companies to maintain their data-mining practices, we could limit them—or shape their practices according to core protected values—by imposing ex post liability. This could be civil, administrative, or even criminal. It could be promoted by fines, for instance, akin to what the GDPR imposes,[143] or it could be reputational and monetary like data breach notifications, which are currently legislated by states.[144] But these forms of regulation, which rest somewhat on deterrence theory, might also prove ineffective for the intended goals.[145]

---

141. As history shows from terms of service agreements, end-user license agreements (EULAs), and privacy policies, most consumers do not bother reading them for two main reasons: these documents are usually long and written in a legal language almost incomprehensible to most people, and consumers today already experience information flooding. *See, e.g.*, Daniel B. Ravicher, *Facilitating Collaborative Software Development: The Enforceability of Mass-Market Public Software Licenses*, 5 VA. J.L. & TECH. 11, 13 (2000); Garry L. Founds, *Shrinkwrap and Clickwrap Agreements: 2B or Not 2B?*, 52 FED. COMM. L.J. 99, 100 (1999); George R. Milne & Mary J. Culnan, *Strategies for Reducing Online Privacy Risks: Why Consumers Read (or Don't Read) Online Privacy Notices*, 18 J. INTERACTIVE MARKETING 15, 20–21 (2004); Joel R. Reidenberg et al., *Privacy Harms and the Effectiveness of the Notice and Choice Framework*, 11 I/S: J.L. & POL'Y FOR INFO. SOC'Y 485, 491 (2015).

142. To name a few reasons, it is usually difficult for individuals to know when their rights were violated, to prove these violations, and to satisfy the injury-in-fact standing requirement under Article III of the Constitution without concrete harm. *See* U.S. CONST. art. III; Lujan v. Defs. of Wildlife, 504 U.S. 555 (1992); Spokeo, Inc. v. Robins, 136 S. Ct. 1540, 1545 (2016) (holding that a plaintiff does not satisfy Article III standing without identifying a concrete harm). It should be noted, however, that some courts ruled that violation of some Acts could constitute injury in fact sufficient to satisfy standing. *See, e.g.*, Matera v. Google Inc., No. 15-CV-04062-LHK, 2016 WL 5339806, at *14 (N.D. Cal. Sept. 23, 2016) (holding that violations of the Wiretap Act and state law constitute injury in fact); *cf.* Hancock v. Urban Outfitters, 830 F.3d 511, 514 (D.C. Cir. 2016) (holding that injury in fact depends also on the type of information which would be sufficient for standing).

143. *See* Commission Regulation 2016/679, 2016 O.J. (L 119) (EU) 1 (repealing Directive 95/46/EC) (General Data Protection Regulation).

144. Data breach notifications usually require some private and government entities to notify individuals of security breaches of information involving personally identifiable information. *See* David Thaw, *The Efficacy of Cybersecurity Regulation*, 30 GA. ST. U. L. REV. 287, 297 (2014).

145. Deterrence theory had been criticized over the years by many scholars. *See, e.g.*, Dan M. Kahan, *The Theory of Value Dilemma: A Critique of the Economic Analysis of Criminal Law*, 1 OHIO ST. J. CRIM. L. 643, 643–47 (2004).

An ex post approach could also take many other forms. Upon communication, companies could try to assess whether a data chunk should be deemed sensitive and retain only pieces of data deemed non-sensitive. In other words, policymakers could impose obligations on private companies to monitor their communications, analyze them, and opt for such a solution if the data are sensitive. Yet, this approach might also be problematic. For instance, it would be difficult to implement in practice because it would require an in-depth, sometimes subjective, analysis of data to determine its sensitivity. But who will decide whether a piece of data is sensitive or not? Should the state delegate this power to quasi-judicial or private entities?[146]

This measure would most likely be taken by computerized systems, as it would be impractical—and not necessarily desirable—to assign individuals to make these decisions, considering that it would be impossible for a human being to perform this task with the necessary accuracy and efficiency.[147] In addition, many companies might not have the capacity to conduct such analyses. Thus imposing strict obligations to review the sensitivity of data on companies might raise the barrier to entry in a market. Here, perhaps, it is even preferable that companies adhere to an "ignorance is bliss" approach, since obliging companies to obtain actual knowledge of the information that is conveyed and stored might defeat the very purpose of safeguarding users' privacy.

Overall, perhaps technology will eventually make the sectoral approach obsolete. Much like technology-sparked discussions on information privacy for a specific cohort (i.e., children) or a specific context (i.e., video rental), IoT might challenge the current perception of what data should be protected, and Congress might eventually add more protections to other types of cohorts or contexts. This might make sense, as the technological changes of the always-on era could be perceived as much more comprehensive in the sense of privacy than those that Congress has regulated over time. But perhaps technology should be viewed as not only the problem, but also the solution for protecting privacy in the always-on era.

---

146. The practice of delegating quasi-judicial powers to intermediaries is, however, not unheard of. The Digital Millennium Copyright Act (DMCA), for example, created a notice-and-takedown regime against copyright infringement and de facto required search engines to "receive requests from copyright owners or their representatives to remove search results that link to allegedly infringing materials." Eldar Haber, *Privatization of the Judiciary*, 40 SEATTLE U. L. REV. 115 (2016). Another example is the so-called right to be forgotten (or right to erasure) in the European Union, which also obliges some intermediaries to delist or even delete data that relate to the right of information privacy under some circumstances. For an overview and criticism of these and other privatization practices, see *id.*

147. *Id.* at 144 (discussing the costs of content reviewers).

The next Part shows how various technological measures could be embedded within IoT devices or services to better protect the values that the sectoral approach sought to protect, offering a toolkit for policymakers and OSPs to embrace a new approach, with or without legal intervention.

IV. REGULATING THE ALWAYS-ON ERA THROUGH TECHNOLOGY

In the always-on era, sensitive data are increasingly collected by unregulated entities under federal laws. Technology in the context of privacy protection, however, is not simply a problem, but may also be a viable solution. Accordingly, this Part discusses potential technological solutions to properly balance data utility with privacy interests and proposes the use of what will be defined as coresets for differential privacy and homomorphic encryption to be embedded in the operation of IoT devices. This Article proposes to use differential privacy ex ante based on the probability of sensitivity, as illustrated in the final Section.

A. TECHNOLOGY AS A SOLUTION

Using technology could enhance privacy protection for users even without abandoning the sectoral approach. In other words, technology might substantially help protect the very same values that it might help infringe.[148] But before discussing specific technological solutions to enhance privacy protection in the always-on era, it is essential to first acknowledge that such protection might depend greatly on the ways they are implemented. As a general framework, this Article advocates the use of an approach termed Privacy by Design (PbD): a "systematic approach to designing any technology that embeds privacy into the underlying specification or architecture."[149] As a concept, PbD could be implemented to help manage various privacy challenges.[150] For example, this concept could be interpreted as calling for structural support for privacy protection and advocating privacy protection by an organization's default mode of operation.[151] As explained later, PbD could be embedded in any technological solution.

---

148. *See, e.g.*, Urs Gasser, *Recoding Privacy Law: Reflections on the Future Relationship Among Law, Technology, and Privacy*, 130 HARV. L. REV. F. 61 (2016).

149. *See* Ira S. Rubinstein, *Regulating Privacy by Design*, 26 BERKELEY TECH. L.J. 1409, 1411–12 (2011).

150. *See* Gasser, *supra* note 148, at 65–66.

151. *See* Ann Cavoukian, *Privacy by Design: The 7 Foundational Principles*, IPC (Jan. 2011), https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf [https://perma.cc/Z94X-NTJK].

This will not be the first instance where technology is suggested—and used—as a solution to protect privacy. This trend began in the 1970s with the development of Privacy-Enhancing Technologies (PETs), designed to be responsive to new information and communication technologies.[152] Many of these technological measures had been suggested or even implemented to protect privacy in the past, but with only modest success.[153] However, PETs could assume many forms, as this Section will show. Generally, privacy protection will adhere to various methods of de-identification of personal data. The ultimate goal will be to preserve the value of data and provide safeguards against identifying users at the same time. These methods could include, inter alia, anonymization and encryption. Here, "identifying users" means that one would not be able to learn from the resulting output (e.g., noisy database or statistics) regarding an individual record of the original database. The formal definition or lack of definition of privacy or the underlying assumptions/ model is a crucial issue that is discussed in the next paragraphs.

We begin with one of the most common techniques in practice, one somewhat infamous in cryptography: data anonymization. Under this method, information in databases could be manipulated to make it intuitively difficult to identify data subjects.[154] Anonymization could be achieved by a variety of techniques (e.g., suppression of data,[155] generalizing identifiers,[156] or providing only aggregate statistics).[157] For example, Congress effectively chose anonymization to regulate healthcare data under HIPAA by specifying eighteen data identifiers whose removal from a dataset would, allegedly at least, protect privacy.[158]

---

152. *See* Gasser, *supra* note 148, at 65.

153. Such technological measures include adblockers, cryptography, and virtual private networks, to name a few. Another method focuses on the output of a query to a given database, meaning that the input itself is not in play, but rather through computation—the output of a query could aid in affording privacy protection. While these measures could play an important role in protecting privacy, they are generally insufficient to grant proper protection to all consumers in the IoT age. *See* Rostow, *supra* note 40, at 694–95; Kobbi Nissim et al., *Bridging the Gap Between Computer Science and Legal Approaches to Privacy*, 31 HARV. J.L. TECH 689, 695–96, 702 (2018).

154. *See* Ohm, *supra* note 61, at 1707–08.

155. Suppression means removing all identifying features from a dataset. *Id.* at 1707.

156. One technique would be suppressing or replacing users' IDs that appear in each record. For example, during the 1990's, America on-line (AOL) collected internet search queries of its users and published them for the research community. To preserve privacy, each user's name was replaced by a random ID number. *See* Karim Z. Oussayef, *Selective Privacy: Facilitating Market-based Solutions to Data Breaches by Standardizing Internet Privacy Policies*, 14 B.U. J. SCI. & TECH. L. 104–05 (2008).

157. *See* Ohm, *supra* note 61, at 1714–16.

158. *See* 45 C.F.R § 164.514(b)(2)(i)–(ii) (2018).

Data anonymization sounds like an almost-perfect solution to protect privacy, but it is not enough. While anonymization and aggregated data could help privacy protection,[159] it does not ensure privacy or protect it sufficiently.[160] Using reidentification or deanonymization methods, an adversary can link anonymized records to auxiliary information and discover the identity of data subjects.[161] This has proven possible by researchers in many instances.[162]

A famous example is Netflix, which publicly released one hundred million anonymized records that revealed how users rated movies. They did so by allowing teams to compete to improve their recommendation algorithm to win the "Netflix Prize."[163] But it was not long before researchers proved how easy it was for an adversary to reidentify many data subjects using merely a smattering of outside knowledge about the subjects' movie-watching preferences. Researchers combined Netflix's published records of movie reviews with other public data, such as IMDB recommendations, that partially matched those records, thereby potentially revealing sensitive data about those

159. *See* Ira S. Rubinstein et al., *Data Mining and Internet Profiling: Emerging Regulatory and Technological Approaches*, 75 U. CHI. L. REV. 261, 266, 268 (2008).

160. *See generally* Ohm, *supra* note 61, Andreas Haeberlen et al., *Differential Privacy Under Fire*, PROC. 20TH USENIX SECURITY SYMP. 1 (Aug. 12, 2011), http://www.cis.upenn.edu/~ahae/papers/fuzz-sec2011.pdf [https://perma.cc/EFD4-CGKS]; Andrew Chin & Anne Klinefelter, *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study*, 90 N.C. L. REV. 1417, 1417–28 (2012).

161. *See* Ohm, *supra* note 61, at 1707–08; Michael Barbaro & Tom Zeller, *A Face is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES (Aug. 9, 2006), http://www.nytimes.com/2006/08/09/technology/09aol.html [https://perma.cc/7MW6-GUGR].

162. *See* Ashwin Machanavajjhala et al., *L-diversity: Privacy Beyond K-anonymity*, 22ND INT'L CONF. ON DATA ENGINEERING IEEE (2006); Josep Domingo-Ferrer & Vicenç Torra, *A Critique of K-Anonymity and Some of Its Enhancements*, THIRD INT'L CONF. ON AVAILABILITY, RELIABILITY & SECURITY IEEE (2008). For further examples, see Latanya Sweeney, *K-Anonymity: A Model for Protecting Privacy*, 10 INT'L J. UNCERTAINTY, FUZZINESS & KNOWLEDGE-BASED SYSS. 557 (2002); Latanya Sweeney, *Weaving Technology and Policy Together to Maintain Confidentiality*, 25 J.L. MED. & ETHICS 98 (1997); Latanya Sweeney, *Simple Demographics Often Identify People Uniquely*, DATA PRIVACY LAB TECHNICAL REP. (2000); Cynthia Dwork, *Differential Privacy*, AUTOMATA, LANGUAGES & PROGRAMMING, 33RD INT'L COLLOQUIUM PROC. PART II 1 (2006); Yakowitz, *supra* note 136, at 3.

163. *See Netflix Prize*, NETFLIX, https://www.netflixprize.com [https://perma.cc/9JAY-THEK] (last visited Jan. 9, 2020) ("The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences."); *see also* James Bennett & Stan Lanning, *The Netflix Prize*, 2007 PROC. KDD CUP & WORKSHOP (2007).

Netflix users.[164] Eventually Netflix settled a class-action lawsuit regarding these potential privacy violations.[165]

Furthermore, de-identification methods like anonymization are notably advancing. One such common approach is a privacy model called k-anonymity.[166] It defines models that provide desiderata with provable guarantees but only under certain threat models (ways that the adversary may attack).

An example privacy mechanism that is used to satisfy k-anonymity is to remove features from each record so that there will be at least k duplications of each record in the data set. The idea is that, as a result, someone possessing the data set will not be able to distinguish among the records in such a cluster yet can still extract data utility from statistics. The main disadvantage of this approach is that we can learn about the user by learning from other records in a user's cluster. For example, each individual in the dataset can easily recognize her own record and thus her cluster, so if all but one person in the cluster band together, they can deduce the remaining person in the cluster. Over the years, heuristics have been suggested to cure this problem,[167] but k-anonymity was nevertheless criticized when researchers proved that k-anonymity and its variants could not preserve privacy in principle and for most basic definitions of the term.[168] Moreover, it might become even less effective for protecting privacy in the IoT context, since the data are usually signals (e.g., audio, video, and GPS) and not strings. For example, it makes sense to remove the last digits

164. More specifically, Netflix offered an award for those that will improve the rating prediction of their users by more than 10%. To do so, they published records of movie ratings from thousands of "anonymized" users. Researchers compared these records with published IMDB records that included the names of reviewers. Since it is very unlikely that a pair of users will give exactly the same rank, even for as little as four movies, it was fairly easy to identify users in Netflix database by comparing them to the IMDB records, often revealing users' sex or political preferences, among other sensitive attributes. *See* Arvind Narayanan & Vitaly Shmatikov, *How to Break Anonymity of the Netflix Prize Dataset*, ARXIV (Oct. 18, 2006), https://arxiv.org/abs/cs/0610105 [https://perma.cc/2SJQ-7MTL]; Ryan Singel, *Netflix Cancels Recommendation Contest after Privacy Lawsuit*, WIRED (Mar. 12, 2010), https://www.wired.com/2010/03/netflix-cancels-contest [https://perma.cc/688X-4ZSC]; Ohm, *supra* note 61, at 1720–21; *see also* Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, PROC. 2008 IEEE SYMP. ON RES. IN SECURITY & PRIVACY 111 (2008).

165. Steve Lohr, *Netflix Cancels Contest Plans and Settles Suit*, N.Y. TIMES BITS BLOG (Mar. 12, 2010, 2:46 PM), http://bits.blogs.nytimes.com/2010/03/12/netflix-cancels-contest-plans-and-settles-suit/ [https://perma.cc/XW5B-7G4K]; Ohm, *supra* note 61, at 1722.

166. *See* Sweeney, *supra* note 162, at 557.

167. *See* Josep Domingo-Ferrer & Vicenç Torra, *supra* note 162 (criticizing systematically others' suggested heuristics).

168. *See, e.g.*, Rolando Trujillo-Rasua & Josep Domingo-Ferrer, *On the Privacy Offered by (K, ⊿)-Anonymity*, 38 IEEE INFO. SYSS. 491 (2013).

of zip codes in order to try to preserve privacy as usually done in k-anonymity; however, it is less clear what to remove from a GPS or a speech signal.

In other words, protecting privacy and innovation with de-identification methods proves a double-edged sword. Only aggressive suppression of data could make reidentification or deanonymization almost impossible, but such suppression would also make the data almost useless.[169] While prohibiting reidentification could seemingly help resolve this puzzle, such a mechanism might be difficult to implement and enforce.[170] When adversaries can easily and legally learn from publicly available information, anonymization methods will not properly advance privacy protection. Thus, in the IoT context, even with data anonymization adversaries might still be able to reveal the identity of the data subject.[171]

We now turn to one of the most common ways to protect privacy: encryption—a field usually related to security.[172] In the context of security, encryption is now ubiquitous for transmitting sensitive data online, such as a credit card number. Even if an unauthorized party saw the communication, it would learn nothing about the transmitted data in cleartext. Encryption is also effectively used for some IoT communications, meaning that an adversary that might view a communication without a decryption key will not be able to extract any data from the content of the message.[173] In some instances, such as in Google Drive, even the stored data might be encrypted, in a way that only the user (not even Google) will possess the secret encryption key.[174]

---

169. *See* Ohm, *supra* note 61, at 1714.

170. *See id.* at 1758.

171. Notably, these de-identification methods are analogous to early cryptography, from simple algorithms such as Caesar cipher that naively add, say, the number three to each letter, to the complicated Enigma machine of World War II whose code was broken during the war by Alan Turing. Like the new de-identification methods, these cryptographic schemes intuitively looked good, but in fact were broken by researchers, sometimes with the help of additional external databases or prior knowledge. In contrast, modern cryptography is based on provable reductions to mathematical problems that are assumed to be too hard to solve in practice and reasonable time using state-of-the-art software and hardware.

172. *See* Hui Suo et al., *Security in the Internet of Things: a Review*, 3 COMPUTER SCI. & ELECTRONICS ENGINEERING (ICCSEE) 650 (2012).

173. Amazon, for instance, declares that it encrypts all communication between the Amazon Echo, the Alexa App, and Amazon servers. *See* Kate O'Flaherty, *How to Secure the Amazon Echo*, FORBES (May 25, 2018, 2:26 PM), https://www.forbes.com/sites/kateoflahertyuk/2018/05/25/amazon-alexa-security-how-secure-are-voice-assistants-and-how-can-you-protect-yourself/#476433cb3734 [https://perma.cc/NB83-MSP4].

174. *See* Darren Quick & Kim-Kwang Raymond Choo, *Google Drive: Forensic Analysis of Data Remnants*, 40 J. NETWORK & COMP. APPLICATIONS 179, 179 (2014). It should be emphasized that even if attackers cannot read the encrypted message, they may still learn meta-data regarding the message (e.g., when it was sent? What is its length? Etc.). Simple possible solutions were suggested in Adi Akavia et al., *Secure search on encrypted data via multi-ring sketch*,

In many cases, however, encryption of IoT data poses problems for innovating and providing services. Many, if not most, IoT devices provide services that rely on data processing, and it is vital to learn from many users' data for machine learning, deep learning, and big data analysis.[175] In this sense, it is vital that the OSPs learn from a private dataset, rather than just store an encrypted version of it. Thus, classic encryption will generally be problematic for IoT data.[176]

To some extent, new forms of encryption methods can help preserve users' privacy while maintaining data utility. One example is homomorphic encryption—a research area in cryptography that aims to solve the problem of outsourcing the computational task without risking privacy.[177] Generally, homomorphic encryption is designed to enable the server or cloud to run computation services without learning anything about the transmitted data in cleartext, by running it on the encrypted data and returning an encrypted result.[178] Hence, unlike standard encryption techniques, homomorphic encryption ensures that only the user possesses the secret key, while computations can be performed on the encrypted IoT data.[179]

2018 ACM CONF. ON COMPUTER & COMM. SECURITY (where the client communicates with the server all the time, possibly using dummy message, in order to hide the time stamp of the real messages).

175. *See, e.g.*, Mohammad Saeid Mahdavinejad et al., *Machine Learning for Internet of Things Data Analysis: a Survey*, 4 DIGITAL COMM. & NETWORKS 161 (2018).

176. For example, in order for Amazon's Alexa to answer a user's question, Amazon must not only obtain the user's voice records, but also process them and return the answer to the user. The processing further requires using Alexa's powerful computation service and accessing all of its databases. *See* Hyunji Chung et al., *Digital Forensic Approaches for Amazon Alexa Ecosystem*, 22 DIGITAL INVESTIGATION 15 (2017).

177. *See* CRAIG GENTRY & DAN BONEH, A FULLY HOMOMORPHIC ENCRYPTION SCHEME 20 (2009).

178. The question whether it is even possible to run any algorithm on encrypted data without knowing the secret key was raised in 1978, within one year of the development of RSA—the first and most common message encryption algorithm. *See* Ronald L. Rivest et al., *On Data Banks and Privacy Homomorphisms*, *in* FOUNDATIONS OF SECURE COMPUTATION 169 (1978). For over thirty years, it was unclear whether a solution, called a fully homomorphic scheme, existed. The first construction was suggested only in 2009 and was considered a major theoretical breakthrough. *See* Craig Gentry, *Fully Homomorphic Encryption Using Ideal Lattices*, 41 ACM SYMP. ON THEORY OF COMPUTING (STOC) 2 (2009).

179. To exemplify, suppose that the IoT device (client) wants to solve a problem or compute f(D) on its data D, where f is the desired function, task, or algorithm. The client encrypts the data D to get its encrypted version [D] and sends it to the cloud. Homomorphic encryption allows the cloud to compute [f(D)], the encrypted version of f(D), using only [D]. It then sends [f(D)] to the IoT device that decrypts [f(D)] using its internal secret key and obtain the result f(D) to perform the client's command.

Today, however, homomorphic encryption is used relatively rarely, as it runs into practical barriers.[180] In the context of privacy for IoT, using this method entails two main disadvantages. First, while homomorphic encryption solves the computational outsourcing problem on the cloud, it does not enable the OSP to learn the transmitted data in cleartext from users' statistics to improve its model, since we cannot learn from encrypted data without having its key. Second, while homomorphic encryption might sound like a good solution in theory, in practice it is known to be unwieldy and unworkable, except for very simple tasks of adding encrypted numbers.[181] In particular, while in theory any algorithm can be applied to the encrypted data, hardly any machine-learning algorithms that can run in this model exist in practice. This makes homomorphic encryption currently unsuitable for many, if not most, IoT services that run machine-learning algorithms.[182]

The potential technological solutions presented in this Section alone are therefore currently ineffective for preserving users' privacy and data utility. Still, technological solutions can be viable if they combine new techniques with at least some of the existing techniques presented in this Section. As the next Section argues, a relatively new approach in computer science could effectively preserve users' confidentiality (to some extent) while keeping data utility at a proper level for the IoT context.

B. DIFFERENTIAL PRIVACY USING CORESETS

Instead of focusing on protecting merely personally identifiable information as in the method of anonymization, this Article suggests focusing on the data subjects themselves.[183] By doing so, this Section proposes a model that can manage the practical and utility issues arising from handling IoT data and preserve provable guarantees regarding users' privacy at the same time. We intend to expand the use of mathematical tools in the context of privacy,[184] while further addressing the core values of the sectoral approach in the always-

180. Wei Wang et al., *Accelerating Fully Homomorphic Encryption Using GPU*, 2012 IEEE CONF. ON HIGH PERFORMANCE EXTREME COMPUTING IEEE 1 (2012).

181. *See id.*

182. Exact running times and performance measures can be found in Miran Kim et al., *Secure Logistic Regression Based on Homomorphic Encryption: Design and Evaluation*, 6 JMIR MED. INFORM. 1, 1–3 (2018).

183. *See* Nissim et al., *supra* note 153, at 687–88.

184. For other suggestions to combine mathematical tools within the notion of privacy protection, see Omar Chowdhury et al., *Privacy Promises That Can Be Kept: A Policy Analysis Method with Application to the HIPAA Privacy Rule*, PROC. 18TH ACM SYMP. ON ACCESS CONTROL MODELS & TECH. 3 (2013); Henry DeYoung et al., *Experiences in the Logical Specification of the HIPAA and GLBA Privacy Laws*, PROC. OF 9TH ACM WORKSHOP ON PRIVACY IN ELECTRONIC SOC'Y (2010).

on era in the following Section. Our model is based on a combination of a few recent techniques in the theory of differential privacy,[185] computational geometry,[186] and homomorphic encryption.[187] The link among these techniques is a modern data summarization technique named coresets (or core-sets), as will be further explained.

Introduced in 2006, differential privacy is a standard that strives to assure that the presence or absence of an individual in a dataset does not make any significant difference to the outcome of any given database query.[188] It mathematically ensures that breaking confidentiality will be limited in probability,[189] and that individuals' data could remain in the database without anyone knowing that it exists.[190] It does so by sanitization of the data (i.e., by adding noise ("blur") to the data) in order to hide information about individual users, while keeping the global statistics, or the ability to construct efficient classifiers from the sanitized data.[191]

Before exemplifying the use of differential privacy in the context of IoT, first it is important to clarify how noise could be introduced. Deciding the level of sanitization or noise to be added to the data requires discussion of two computation models and communication protocols: centralized or local. Under a centralized model, the OSP collects the data from its users and is also responsible for adding the noise.[192] The original data must be deleted or at least not be used by the OSP prior to adding the noise. The data will then be sanitized with noise, and the learning algorithms will be fed the "sanitized

185. For more on differential privacy, see Cynthia Dwork, *Differential Privacy: A Survey of Results*, *in* THEORY AND APPLICATIONS OF MODELS OF COMPUTATION 1 (Manindra Agrawal et al. eds., 2008).

186. *See* Dan Feldman & Michael Langberg, *A Unified Framework for Approximating and Clustering Data*, PROC. 43D ANN. ACM SYMP. ON THEORY COMPUTING 569, 569–71 (2011).

187. *See* Adi Akavia et al., *Secure Search on the Cloud via Coresets and Sketches*, ARXIV (Aug. 19, 2017), https://arxiv.org/abs/1708.05811 [https://perma.cc/Z4C5-HAF7].

188. *See* Cynthia Dwork et al., *Calibrating Noise to Sensitivity*, PRIVATE DATA ANALYSIS, PROC. 3RD CONF. THEORY OF CRYPTOGRAPHY 265 (2006); Chin & Klinefelter, *supra* note 160, at 1427 (citing Cynthia Dwork, *A Firm Foundation for Private Data Analysis*, COMM. ASS'N FOR COMPUTING MACHINERY 86, 91 (2011)). For more on differential privacy, see Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1139–40 (2013); Ohm, *supra* note 61, at 1756; Chin & Klinefelter, *supra* note 160, at 1452–54; Jane Bambauer et al., *Fool's Gold: An Illustrated Critique of Differential Privacy*, 16 VAND. J. ENT. & TECH. L. 701, 712–17 (2014).

189. *See* Ohm, *supra* note 61, at 1756.

190. *See* Chin & Klinefelter, *supra* note 160, at 1430.

191. *See* Shuchi Chawla et al., *Toward Privacy in Public Databases*, 2 THEORY OF CRYPTOGRAPHY CONF. (2005).

192. *See, e.g.*, Xi Xiao et al., *CenLocShare: a Centralized Privacy-preserving Location-sharing System for Mobile Online Social Networks*, 86 FUTURE GENERATION COMPUTING SYSS. 863 (2018).

data." The centralized model is often used as a method for allowing data use by third parties or for spreading data in different OSP departments, thus lowering the risk of information leakage to competitors by employees.[193] But this model is weakened by its reliance on trust and efficiency. The method relies on trusting the OSP to actively sanitize data and on sanitization by OSPs, proving too late for some users. Data from IoT devices might be hacked, stolen, lost, or just poorly sanitized.[194]

The alternative model is local, wherein the IoT client does not share its raw data with the OSP.[195] Instead, the sanitization is done ex ante on the client's side. Only the sanitized dataset is sent to the OSP. As a result, trusting the OSP and preventing data leakage prior to sanitization are not a challenge as long as the sanitization is performed properly. The disadvantage of the local model is that the per-user raised noise level is significantly greater than in the centralized approach, where small added noise suffices to blur the original statistics. In addition, in many instances locally added noise is still too low to preserve users' privacy.[196]

Unlike previous techniques, such as k-anonymity, which also rely on mechanisms such as adding noise or hiding data, differential privacy suggests a very strong definition of privacy that is resistant to external databases that the adversary may have. An algorithm is differentially private only if it provably meets this definition. Such an algorithm is unlike de-identification techniques, with which a sanitized database has noise added per record (and the adversary

193. For example, some argue that Facebook uses this technique to publish click rates to its ad publishers. *See* Yehuda Lindell & Eran Omri, *A Practical Application of Differential Privacy to Personalized Online Advertising*, 2011 IACR CRYPTOLOGY EPRINT ARCHIVE 152 (2011). This method is also common in governments' Bureau of Statistics in order to publicly share their collected data. *See* Boaz Barak et al., *Privacy, Accuracy, and Consistency Too: a Holistic Solution to Contingency Table Release*, PROC. 26TH ACM SIGMOD-SIGACT-SIGART SYMP. ON PRINCIPLES DATABASE SYSS. (2007).

194. *See, e.g.*, Ryan Singel, *Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims*, WIRED (Dec. 17, 2009), https://www.wired.com/2009/12/netflix-privacy-lawsuit [https://perma.cc/BP47-FGS3]; Narayanan & Shmatikov, *supra* note 164, at 6–10.

195. For more on the local model, see generally Peter Kairouz et al., *Extremal Mechanisms for Local Differential Privacy*, 17 J. MACHINE LEARNING RES. 1 (2016).

196. For example, instead of collecting GPS data from its users and adding noise to it (i.e., centralized privacy), Apple added sanitization mechanisms on its smartphones, so some type of noise is added to the GPS samples before transmitting them to Apple from the user's smartphone. However, researchers that reverse engineered this protocol claimed that the amount of noise added is far too small to preserve users' privacy. *See* Andy Greenberg, *How One of Apple's Key Privacy Safeguards Falls Short*, WIRED (Sept. 15, 2017, 09:28 AM), https://www.wired.com/story/apple-differential-privacy-shortcomings [https://perma.cc/LAN5-K7LC]; Jun Tang et al., *Privacy Loss in Apple's Implementation of Differential Privacy on macOS 10.12.*, ARXIV (Sept. 11, 2017), https://arxiv.org/pdf/1709.02753.pdf [https://perma.cc/4R58-KH6V]. *See generally* Chin & Klinefelter, *supra* note 160.

can tell, for example, if another user has been added since the previous version of the database). Instead, a differentially private algorithm completely replaces the database with a new database containing "global" noise.

The problem, however, is that most of the literature in modern computer science that discusses privacy, including k-anonymity and differential privacy, does not generally fit the IoT model. In particular, as was the case with the Netflix Prize, the literature assumes that the OSP holds the complete original (non-sanitized) database of its users. When that is the case, privacy issues are assumed to arise only once the OSP reveals its user data to a third party (e.g., when Google sells or disclose parts of its database to advertisers) or at least reveals a derivation of the data, such as classifiers or statistics, that may leak information about individuals.

By contrast, in the foregoing Sections we assume that there are *many* users that send their private data to the OSP and wish to preserve their privacy. While in principle the OSP might itself add noise to the collected records—that is, after collecting and before using them—this is too late if the users do not trust the OSP. The hidden and natural implication is that the noise should be added *locally* on the user's side *before* even reaching the OSP. This communication model for privacy was suggested relatively recently and requires much more noise to be added than the centralized model requires.

To understand how differential privacy could help in the always-on era, we begin with an example: baby diapers. Suppose a company that sells diapers asks how many Amazon Echo users have a baby at home. The motivation may be to decide where to place their ads or perhaps to use the device itself for marketing purposes. This can be done, for example, by listening for a baby crying at some point in time or analyzing the voice after a conversion. If $x_i = 1$ when the ith client has a baby at home and $x_i = 0$ when not, the answer when there are n clients is the sum: $S = x_1 + x_2 + \cdots + x_n$. Suppose that the diapers company already knows some of these values (e.g., the sum of the first n-1 numbers in this equation), through either relying on another database or computing this number on the day before the last *n*th client joined.

Within the diapers example, if Amazon publishes this number, and the diapers company knows the sum of the first n-1 clients, they will be able to compute the value $x_n = S - x_1 - \cdots - x_{n-1}$ of the last client. That is, they will be able to compute whether the *n*th client has a baby at home. To avoid this, Amazon could compute the sum S and add a little noise before presenting the noisy value of S to the diapers company. Even if the diapers company knows all the values of x_i except one, as long as sufficient noise was added, the company will not be able to extract x_i from the noisy value of S. Thus, they will not be able to compute whether the *n*th client has a baby at home.

Some scholars claim Facebook already uses this technique for sharing information about the number of its users' clicks for third-party sponsors.[197]

The main challenge in using differential privacy is knowing how to add noise to the data that is sufficiently large to preserve the individual's privacy, but sufficiently small to allow a good approximation of useful statistics. For example, if a random number is added to a given sum of numbers from a Laplacian[198] distribution with zero mean and scale of roughly $1/\varepsilon$, where ε is a number usually between 0 and 1, then the adversary that receives S will (depending on the value of ε) be unable to determine whether any specific x_i is 0 or 1, even if the adversary knows that this particular algorithm outputted S.

More precisely, the probabilities that $x_i = 0$ and $x_i = 1$ given S are approximately the same, up to an additive error of ε. In other words, whatever the adversaries know or wish to know regarding a specific value x_i, and whatever external database or knowledge they already have, they will not be able to learn about x_i merely because it was part of the original input. Formally, the output of the randomized algorithm that computes the approximation of S has the same distribution (up to ε additive factor) if we change a single value. True, a dummy algorithm that outputs a random number will also satisfy this privacy guarantee. However, the above private algorithm is more desirable because it is efficient: with high probability, depending on ε, it gives a good, provable approximation of S. This property of allowing approximation is called the utility of the algorithm.

In the diapers example, we assumed that Amazon computes S in a centralized private fashion. But a locally private version for the above solution is also possible. If we do not want the actual values of x_i to get to Amazon via its Echo device in the first place, each ith device should add its own noise to its value x_i. A common approach that has provable guarantees is to send the "wrong" value with some fixed probability. For example, an algorithm might send as its vote the real binary value x_i with probability 0.75, and otherwise send the "wrong" value $1 - x_i$.[199] The privacy of each user is preserved in the sense that with 0.25 probability (a 25% chance), x_i is not the real value, while

197. *See generally* Chin & Klinefelter, *supra* note 160.

198. The Laplacian distribution is used since it is proportional to $\exp(|x|)$, which yields the desired property $\exp(|x|+\varepsilon)/\exp(|x|)=\exp(\varepsilon)$. This property does not hold for, for example, the Gaussian distribution that is proportional to $\exp(x^2)$. For further detail, see sources cited *supra* note 188; Dan Feldman et al., *Private Coresets*, PROC. 41ST ANN. ACM SYMP. ON THEORY OF COMPUTING (2009).

199. *See generally* Stanley L. Warner, *Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias*, 60 J. AM. STAT. ASSOC. 63 (1965).

the utility is preserved for a sufficiently large number of users: for example, if the approximated value of S is $/4$, then probably most of the users sent 0, and $n/4$ is due to the noise. On the other hand, if $S = \frac{3n}{4}$, the real value of S is close to n. In both of these cases, there is a "doubt" of probability 0.25 whether each user's vote was her real value or the opposite. This is regardless of her real value, or the real values of the other people in the group. Note that the expected error of 0.25 in this local privacy model is by order of magnitude larger than the expected error in the previous centralized model of $\frac{1}{n}$, which also decreases with the number of users.

Notably, the use of differential privacy as a solution for privacy has also been subjected to criticism in various respects. Scholars argue that introducing noise is limited in value for a few reasons: the use of noisy data might yield inaccurate outcomes; it might require complex and costly calculations; and it could cause chaos in database systems.[200] It will also not apply to data already collected. However, while acknowledging potential shortcomings, some scholars have recently considered differential privacy as a potential solution to protect educational privacy from a legal perspective.[201]

Indeed, one of the main challenges in differential privacy is to explain the intuition behind its guarantee of privacy. A common explanation is that a differential privacy algorithm is private in the sense that an adversary can only obtain information that can be learned *whether you participate in the database or not*. In other words, the algorithm and its output are insensitive to any single user. For instance, given a sanitized database of patients, we may conclude that heavy smoking may cause cancer. So, if we see someone smoking, we can infer that she may have a higher probability of getting cancer. While we learned *something* about this individual from the database, it was not because she participated in the database; indeed, the database may not have contained any information specific to her at all. If her specific information were added to the database, we would learn nothing new about her. Thus, in that sense, her privacy is preserved. In contrast, this property does not hold in alternatives with no provable guarantees, such as in the Netflix case. In those cases, we can often learn information about an individual that we would never have known

200. *See* Arvind Narayanan & Vitaly Shmatikov, *Privacy and Security Myths and Fallacies of "Personally Identifiable Information"*, COMM. ASS'N FOR COMPUTING MACHINERY 24, 26 (2010), http://www.cs.utexas.edu/~shmat/shmat_cacm10.pdf [https://perma.cc/YUG4-RTXH]; Bambauer et al., *supra* note 188, at 704; Ohm, *supra* note 61, at 1757.

201. *See generally* Nissim et al., *supra* note 153 ("[D]ifferential privacy satisfies a large class of reasonable interpretations of the FERPA privacy standard."). Notably, the scholars argue that FERPA and differential privacy were used to illustrate an application of their approach and that it "may be developed over time and applied, with potential modifications, to bridge between technologies other than differential privacy and privacy laws other than FERPA." *Id.*

but for her specific information being included in the database. Admittedly, this intuition about differential privacy can be difficult to grasp.

Another challenge of differential privacy is applying it to more complex calculations. The diapers example involved only simple calculations. However, in applications such as machine learning, the input is not just a binary number, but a long database record of each user. Moreover, the data emitted is not from mere summation, but from more complicated functions such as neural networks or logistic regression. More generally, differentially private algorithms traditionally use impractical models, such as Curator,[202] and solve very specific, and arguably artificial, theoretical problems with guarantees that engineers and lawyers may understand significantly less than the other methods.[203]

One final challenge is practical and arises from utility challenges. In many existing protocols, such as Curator, the user is limited to asking a small number of very specific questions. However, while data scientists and learning algorithms may handle added noise, they usually need to learn a single sanitized database for all their queries. Even when sanitized databases are used today, they are useful for a specific family of problems.

To overcome the challenges of differential privacy, we propose combining it with other techniques, most importantly with the notion of coresets: a small representation of the data, such that querying the coreset will yield a provably small approximation to the original data. In particular, solving an optimization problem or running a learning algorithm on the coreset will yield a near optimal

---

202. Curator is a model of computation. In this model, there is a specific service (e.g., a web interface) with access to the original (non-noisy) data of users that gives noisy answers to given questions. Usually, these questions are restricted to a specific type. This service—the curator—may operate either on the end user's side and answer queries of the company, or between the company that stores the original data and its third-party clients. The curator can usually answer specific types of statistical questions regarding the data with some additional noise. Since each answer admits some small privacy leakage, once a specific number of questions have been asked, the curator refuses to answer further questions in order to avoid too much leakage. The curator model is more common in academic papers than in practice: while it can provide provable privacy guarantees, data scientists are accustomed to working with databases (noisy or otherwise) rather than such answering services. Moreover, machine-learning algorithms are usually applied and trained on data sets, and it is unclear how to use curators with these methods. For more on the curator model, see CYNTHIA DWORK & AARON ROTH, THE ALGORITHMIC FOUNDATIONS OF DIFFERENTIAL PRIVACY 211–407 (2014).

203. This relates to an academic debate that has serious implications for the industry and IoT privacy. In particular, it implicates what it means to preserve the privacy of a user and how to achieve it. *See* Bambauer et al., *supra* note 188, at 712–17; Nancy Victor et al., *Privacy Models for Big Data: A Survey*, 3 INT'L J. BIG DATA INTELLIGENCE 61, 61–65 (2016); George Danezis & Seda Gürses, *A Critical Review of 10 Years of Privacy Technology*, 2010 PROC. SURVEILLANCE CULTURES: A GLOBAL SURVEILLANCE SOC'Y 1 (2010).

solution on the original data.[204] Unlike other forms of lossy compression, like MP4 or JPEG compression, coresets are considered lossy compression for specific optimization problems or statistics to be applied on the data, rather than generic compression of the data itself. Coresets have been suggested in recent years to resolve key problems in machine learning, and a single coreset may be the union of multiple coresets to solve the numerous corresponding problems.[205] By means of a technique called sup-sampling (or non-uniform sampling), a single coreset may handle many problems by uniting coresets of the database for these problems.[206]

Coresets are especially useful in the context of IoT and Big Data in general, since they usually possess an important property: the union of a pair of coresets yields a coreset for the union of the underlying data.[207] This implies that one can compute the coreset on streaming IoT data by compressing small batches of data that arrive on the fly and recompress them. So, at any given moment, we can have a small coreset for all the seen streamed data and thus can apply existing (possibly inefficient) algorithms to the small data. Similarly, the IoT service provider can easily compute coresets on the cloud by computing a coreset in each machine on its own streaming data. Then, a main server can collect the coresets for all the machines and solve the optimization problem on it, possibly after an additional final compression.[208]

---

204. *See* Feldman & Langberg, *supra* note 186.

205. *See* Artem Barger & Dan Feldman, *k-Means for Streaming and Distributed Big Sparse Data*, PROC. 2016 SIAM INT'L CONF. ON DATA MINING, SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS 1–2 (2016); Dan Feldman et al., *Coresets for Vector Summarization with Applications to Network Graphs*, INT'L CONF. ON MACHINE LEARNING (2017).

206. *See* Michael Langberg & Leonard J. Schulman, *Universal ε-approximators for Integrals*, PROC. 21ST ANN. ACM-SIAM SYMP. ON DISCRETE ALGORITHMS SOC'Y INDUS. & APPLIED MATHEMATICS (2010). Even without knowing how to compute a coreset for a given problem or classifier, a coreset for a related problem may suffice. In practice, a coreset generally yields a good approximation for a problem it was not designed to solve, since intuitively, a representative point for one problem is also a good representation for the other problem. A coreset for an optimization problem is usually more general in the sense that it usually can approximate queries of a certain type and not just solve the relevant optimization problem. *See* Rohan Paul et al., *Visual Precis Generation Using Coresets*, 2014 IEEE INT'L CONF. ON ROBOTICS & AUTOMATION (2014).

207. *See* Piotr Indyk et al., *Composable Core-sets for Diversity and Coverage Maximization, in* PROC. 33RD ACM SIGMOD-SIGACT-SIGART SYMP. ON PRINCIPLES DATABASE SYSS. (2014).

208. Indeed, coresets for many fundamental problems in machine learning, including experimental results, appeared during the recent decade in machine learning conferences. *See, e.g.*, Dan Feldman & Tamir Tassa, *More Constraints, Smaller Coresets: Constrained Matrix Approximation of Sparse Big Data*, PROC. 21TH ACM SIGKDD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING (2015); Mario Lucic et al., *Training Gaussian Mixture Models at Scale via Coresets*, 18 J. MACHINE LEARNING RES. 5885 (2017).

The main challenge in the research of coreset construction is thus to prove that, for any possible input set, we can compute a small coreset whose approximation error is small, with a good tradeoff between the approximation error (say ε) and size of the coreset (say $1/\varepsilon$). But how can coresets, which enable efficient compression of the data, help us solve the utility issues of differential privacy? In principle, there is no linkage between the ability of data compression, as in the coresets above, and the ability to add a small amount of noise that will preserve the desired approximation error while maintaining privacy—we can always reduce a sanitized database, using coresets to reduce its size, without losing more privacy. However, a perhaps surprising theorem forges a link between the two: if we have a (non-private) small coreset for a problem, we also can have a sanitized database (called a private coreset) where the size of roughly $1/\epsilon$ of the small coreset turns into the additive error (noise) of a similar order.[209] That is, a (non-private) coreset of small size implies a (not necessarily small) private coreset that is computed via a differential ϵ-private coreset.

Different from the Curator model, such a sanitized database can be queried unlimited times without further information leakage once the private coreset is computed from the raw data. That is, the existence of a small coreset implies a sanitized database that preserves the desired statistics (in terms of utility) and also preserves privacy. This theorem is very promising, (e.g., for machine learning in IoT), since dozens of coresets for main problems are already known. Unfortunately, the proof of the above theorem is not constructive in the sense that the computation time of this generic coreset reduction is impractical. How to implement it efficiently is still an open question. Instead, specific private coresets have been suggested in recent years for specific problems.[210]

Hence, private coresets may be used to obtain a single sanitized database, unlimitedly applicable to many machine-learning algorithms with no additional noise. Now the problem remains of computing a private coreset with little added noise (as in the centralized model) while using the localized model for IoT applications. Recall that the main advantage of centralized models compared to local ones is that less added noise is required, and the main disadvantage is that the company has the users' original (non-noisy) data, whereas in local privacy the user sends only noisy data. Indeed, while private

209. *See* Dan Feldman et al., *Private Coresets*, PROC. 41ST ANN. ACM SYMP. ON THEORY COMPUTING (2009).

210. *See* Dan Feldman et al., *Coresets for Differentially Private K-Means Clustering and Applications to Privacy in Mobile Sensor Networks*, 2017 16TH ACM/IEEE INT'L CONF. ON IEEE 3 (2017).

coresets may be computed on the client's side (as in local privacy) or on the server's side (as in centralized privacy), the following technique may allow us to get the small error of centralized privacy, while preserving the client's privacy as in local privacy. To that end, we suggest computing private coresets using homomorphic encryption.

More precisely, the Homomorphic Encryption Coreset (CHE)[211] is a modern tool that may resolve this conflict, namely, to get a small error without letting the company access the original raw IoT data. We denote by D the database of users' data, and by *sanitized*(D) its sanitized version (private coreset). In the traditional centralized model, users send their records of raw data to the OSP, which maintains the database D and then computes its sanitized version *sanitized*(D) that can be used for publishing or learning without sacrificing privacy. The main challenge is learning how the OSP can compute the private coreset *sanitized*(D) without having access to the original database D.

We suggest using homomorphic encryption to compute differential private coresets as follows: instead of sending their original records, or noise records, the client or clients send an encrypted version of their records (but without noise) [D] of, for example, GPS or an Echo's data D to the OSP. The OSP adds noise to the data, as in the centralized model. However, this is done on the encrypted version of the data so there is no privacy loss at all. The result is an encrypted private coreset [*sanitized*(D)]. Now, this is a private coreset that can be exposed to the OSP, but it is still encrypted. At this point the OSP sends the data back to the clients' IoT device that uses its secret key to decrypt [*sanitized*(D)] and obtain the private (non-encrypted) coreset. This sanitized dataset *sanitized*(D) is sent back to the server, which can use it (e.g., to improve the machine-learning results for other users as well). Under this proposition, only a small amount of noise has been added by the server as in centralized privacy, and still the server has never seen the complete original data. More generally, this problem can be applied to data from multiple users where each user has its own key and the sanitized database is computed for all of them.[212] This is how private coresets via homomorphic encryption can make differential privacy more practical, without losing its theoretical guarantees.

Gaps and many handling problems in IoT still await private solutions, on both the theoretical and practical sides of computations, as well as through laws and regulations. For example, the algorithms for computing sanitized

211. *See generally* Adi Akavia et al., *Secure Search via Multi-Ring Fully Homomorphic Encryption*, 25TH ACM CONF. ON COMPUTER & COMM. SECURITY (2018).

212. *See* Adriana López-Alt et al., *On-the-Fly Multiparty Computation on the Cloud via Multikey Fully Homomorphic Encryption*, PROC. 44TH ANN. ACM SYMP. ON THEORY COMPUTING (2012).

databases are usually applied to static database records. However, IoT is made of streaming data that grow larger with time. There are very few results for handling such streaming data privately, especially without introducing too large an additive noise. Similarly, for most problems it is not clear how to compute the sanitized database in parallel on distributed data (e.g., cloud or smartphones), unless we use local privacy. Private coresets may help in handling these issues because there are simple reductions that show how, given a coreset for a set of models, we can compute it on the streaming and distributed model. The main idea is that two coresets can be computed independently on distributed machines such as the cloud, or different subsets of streaming data. Then they can be merged and reduced again on each machine or device.

Our proposed model thus requires further discussion on when to add noise to IoT devices, and what level of noise will ultimately preserve privacy and remain useful for OSPs. Too much noise will make the data unusable, while too little will defeat the purpose of preserving privacy. Accordingly, the next Section proposes a theoretical ex ante approach which strives to protect privacy by the probability of gathering sensitive data, which will depend on various factors.

C. MEASURING NOISE VIA THE PROBABILITY OF SENSITIVITY

To date, scholars have only limitedly applied differential privacy in the context of sectoral privacy protection by, for example, offering differential privacy to satisfy the requirements of a particular legal standard of privacy (FERPA in this instance).[213] Our intention is to broaden this innovative argument. Using the concept of differential privacy, combined with other mathematical models, we offer an analytic framework for any policymaker—taking the U.S. approach to privacy—to protect privacy while still acknowledging the value of data. Further, the level of noise added to the model could be measured—at least to some extent—on the potential sensitivity of the data, depending on various factors related to the IoT in question.

It is generally difficult to define when data become sensitive, although a few scholars have attempted to do so.[214] Evaluating the probability that data will be sensitive ex ante—under the sensitive categories that Congress has set—is even more ambitious. It is generally an almost impossible task to accomplish. Sectoral privacy, however, is implemented almost precisely by this ambitious method. It is regulated through the notion that with some entities,

213. *See generally* Nissim et al., *supra* note 153.

214. *See, e.g.*, ÉLOÏSE GRATTON, UNDERSTANDING PERSONAL INFORMATION: MANAGING PRIVACY RISKS (2013); Ohm, *supra* note 1, at 1733–34.

and in some contexts, there is increased probability that sensitive data will be shared digitally, as is the case in medical, educational, or financial institutions. This measure could be also implemented—perhaps even more accurately—in the always-on era. Thus, without belittling the important scholarly debate on data sensitivity, this Article focuses on the current categories of sensitive information, as reflected in the federal statutes of sectoral privacy: financial data, health information, education records, children's data, and consumer data. As shown below, considering various factors that relate to IoT, and the nature of the use of these devices, could aid in assessing such probability.

The factors to consider when assessing the probability of IoT devices gathering sensitive data depend on various potential characteristics. As explained below, the factors we suggest include the device's architecture, its sensors, its physical location, the nature of gathered data, and the age of its potential users. The probability of infringing on information privacy should be evaluated through the aggregation of these factors: that is, prior to any query from the database. But this does not mean that these factors should not be reevaluated continuously. On the contrary, we encourage such reevaluation, followed by proper modifications and adaptations. Additionally, these factors might greatly vary depending on their users' input, implemented with the use of the device.

We begin with the architecture. As previously noted, not all IoT devices operate alike. Some might have to be turned on manually to begin their data collection (e.g., the smart connected toy Hello Barbie). Others operate in an "always-ready" mode like Amazon Echo or Google Home, meaning that they await their trigger phrase prior to any data collection or retention. Finally, we have the devices that are "always on" (i.e., that constantly collect and transmit data, like Fitbit).[215]

The architecture of the device could greatly influence the probability of collecting sensitive data. Clearly, and without considering other factors like the types of data collected, devices that constantly collect data will have greater probability of collecting sensitive data than always-ready devices. Consequently, in many instances always-ready devices could have higher probability of collecting sensitive data than those operated manually, simply due to their architecture (i.e., it is much more convenient for many individuals to use them so they could be more frequently used). We suggest that OSPs

215. Fitbit is a fitness tracker that monitors steps and could provide insights on, inter alia, an individual's heart rate or quality of sleep. *See* Andrew Hilts et al., *Every Step You Fake: A Comparative Analysis of Fitness Tracker Privacy and Security*, OPEN EFFECT REP. 3–6 (2016), https://openeffect.ca/reports/Every_Step_You_Fake.pdf [https://perma.cc/89PV-Z5TF].

might be required to differentiate three probability levels, depending on architecture: manual, always-ready, and always-on.

The second factor to be considered is the sensors of the device. They may vary greatly, but we can still further divide them into five categories: sensors that measure the environment (e.g., temperature or air quality), that measure human activity (e.g., movements, location, and heartrate), that capture written communication, that capture oral communication, and that capture visual communication (cameras). The types of sensor could greatly affect the probability of gathering sensitive data. In this regard, every type of sensor could be given a numerical representation related to the value of ε.

The third factor is the device's physical location. Some devices are more portable than others. Some might be placed in locations that may have a greater probability of gathering sensitive data than others. As for the first argument, devices like smart refrigerators or smart washing machines will most likely not change their placement much, while mobile health ("mHealth") wearables, like Fitbit, are more likely to be on the move. Even smart personal assistants can be moved more easily than smart refrigerators or washing machines, hence could be placed anywhere that has connectivity to both external power and the internet.

The placement of such devices could affect the sensitivity of data gathered as some locations could impact the type of data conveyed. For instance, a house's bedroom could convey data on sexual activity more than the kitchen area. Placing an Amazon Echo in your living room might not be the same as placing a similar device in your office. However, this concern must be evaluated with respect to sectoral privacy. It is extremely difficult to determine the link between placement and sensitivity in general, so instead we merely suggest considering two broad categories of devices: wearable and not wearable. A wearable device, especially one constantly worn, is more likely to gather sensitive data simply due to its ability to collect sensitive data directly from an individual more easily.

The fourth factor to be considered is the nature of the gathered data. Some IoT do not gather any sensitive data at all, or at least have a low probability of gathering such data. For instance, if a smart refrigerator knows which food or drinks it contains, and perhaps even consumption habits, it has low probability of collecting sensitive data, if any at all. Other IoT devices have a higher probability of collecting such sensitive data. While Amazon Echo is not generally likely to gather health information in the course of its communications, as its business model does not depend on health data per se, it might do so upon communicating with it, and it could gather consumer-sensitive data if used for purchasing. A smart TV could also fit this category,

as it might be aware of your watching habits and might also capture sensitive communication. Finally, we have the devices that by default collect sensitive data. Health wearables, for example, depend on health data; hence the gathering of such data is almost certain. Thus, the nature of gathered data will depend on the three-fold categorization of their potential nature: low, high, and certain.

The fifth and final factor to be considered is the user's age. As federal law generally protects children younger than thirteen in some circumstances,[216] we must evaluate the probability of having an IoT device gather data from this cohort. Thus, this factor divides devices into two main categories: devices that are, and that are not, targeted at children younger than thirteen. For the devices that do target children, like smart connected toys or kids' wearables, the probability of gathering sensitive data is absolute. This is the easy case, as COPPA applies to the OSPs of these devices which must ensure that they comply with its regulations.[217] This category is thus excluded from this factor, as it is already implemented and labeled "certain" through the fourth factor, namely the nature of gathered data. The second sub-category is more challenging and will depend, inter alia, on users' inputs. When configuring the device, users will be obliged to answer various questions that will help determine the probability that children's data will be obtained. So, if a user operates an Amazon Echo device in his or her living room, having children aged under thirteen in the household will increase the probability of gathering such data. This probability will change depending on the number of children in the household and their cognitive abilities, among other potential factors.

The probability of sensitivity can be calculated through each of these five factors, and perhaps mostly through their correct combination. This calculation will involve an ex ante evaluation of the IoT device in question, along with input from users (e.g., information about whether there are children present in a household) that will allow fine-tuning of such an evaluation. Upon evaluation of these factors, and perhaps others, OSPs could translate them into a relative ordering of whether the risk is high, medium, or low.[218] Such probability could be implemented through the mathematical models we have suggested, thereby adding noise only to the IoT devices that present higher probability of sensitivity without an ex post evaluation of the data or the data

---

216. *See* 16 C.F.R. § 312.2 (2018); 15 U.S.C. §§ 6501(1), 6502, 6501(8) (2012).

217. *See Children's Online Privacy Protection Rule: A Six-Step Compliance Plan for Your Business*, FED. TRADE COMM'N (June 2017), https://www.ftc.gov/tips-advice/business-center/guidance/childrens-online-privacy-protection-rule-six-step-compliance [https://perma.cc/276M-9AH2] (last visited Feb. 10, 2019). For more on smart connected toys and COPPA, see generally Haber, *supra* note 95.

218. *See* Ohm, *supra* note 61, at 1765.

subject. In other words, IoT devices could measure, to some extent, the probability of sensitivity, give it a numerical representation, and add noise to the IoT device according to such privacy risk assessment.

We turn to briefly explain how the value of ε can be adjusted to properly balance data utility and privacy.[219] This Article suggests defining the value of ε based on the probability of sensitivity. IoT devices with a high probability of gathering sensitive data—like Fitbit, which acquires sensitive health data—should add more noise to any gathered data. Meanwhile, IoT devices with a low probability of gathering sensitive data, like smart refrigerators, should add less noise—or none at all—depending on the privacy risk assessment. For example, an Amazon Echo in a household with children under the age of thirteen should add more noise to its data than an Amazon Echo in a household without children, all other things being equal. Essentially, any interaction with IoT technology will require an ex ante evaluation of the probability of sensitivity, followed by adding sufficient noise to the device's operation.

At this point it is essential to underline some caveats. First, our model is built on the current values embedded in the federal sectoral approach. It thus excludes, inter alia, state legislation that might also be relevant for privacy protection. It is also only natural that the perception of sensitivity of data change with technology and potential social changes. Thus, as previously mentioned, this model must be constantly challenged and recalibrated when necessary.

Second, our quantification of the level of data sensitivity could be viewed as somewhat arbitrary. In that regard, our intention is rather modest. We strive to show mostly how data probability could be assessed and used by differential privacy, but we make no binding statements regarding the actual values linked to data or context. These values could be challenged and changed by scholars or policymakers when necessary. Moreover, as previously mentioned, this model must also be adaptive, and obviously also include other types of sensitive data that must always be reevaluated in light of potential technological or social changes. Accordingly, any mechanism of probability is flexible and

219. Calibrating the privacy level or the added noise to the data is related to the value of the privacy parameter ε defined in the context of differential privacy. Calibrating ε depends on the specific application and type of data. When the data can be visualized, like GPS data for example, it may be computed interactively in a graphic way. Such a solution has been suggested, where the user visually sees how the data change in real-time while changing ε via a slide-bar on a graphic user interface. When the data look sufficiently noisy, and the user sees that the secret data cannot be extracted intuitively, the current value of ε is chosen. *See generally* Adi Akavia et al., *supra* note 211.

could be fine-tuned with time, especially if a specific form of technology is found able to collect more sensitive data than the model anticipated.

Third, if policymakers impose obligations on industries because of such an evaluation, the obligations might be detrimental if they are broader than intended or negatively affect the data's utility. This could lead to what is known as the principle of parsimony, meaning that taking broader action under uncertainty might have negative consequences.[220] The negative consequences of using differential privacy, however, are not worrisome. Concededly, some data might become less valuable for industries subject to overbroad regulation. But this potential drawback should be balanced against the benefits of using such a model. Thus, while sometimes differential privacy might be used under conditions of uncertainty, its impact on the quality of the data is not substantial. Essentially, it retraces the trade-off between privacy and utility.

Finally, our technological solution is likely to be combined with the modality of the law. Lacking external incentives for information protection, market actors' self-regulation is bound to fail.[221] There must be some form of incentive for companies to adhere to these requirements. This could be achieved, for example, by obliging private companies to implement these technological measures ex ante in order to begin operating (e.g., by requiring licenses) or ex post (by imposing high fines for noncompliance or data breaches). It could also be achieved by granting a safe harbor from liability lawsuits on the fulfillment of these standards, which will be treated as evidence of compliance vis-à-vis liability or even combining the modalities of social norms and the market to drive consumers to demand that these companies protect their privacy better.

With these caveats in mind, the main purpose of this Article is not to provide a definitive formula that will apply perfectly in every context, not to mention that is well near impossible to achieve. Its intention is to introduce a new mechanism that combines the notions of privacy perception with differential privacy, thus providing a relative form of privacy. This form of privacy is not only a practical means for privacy protection, but it also broadens the discussion on the use of technology to meet new challenges better. Without adhering to such methods, regulating privacy in the always-on era by the sectoral approach will defeat many of the purposes behind such forms of legislation, and will ultimately fail to properly protect individuals' privacy.

---

220. *See* Schwartz, *supra* note 74, at 923 (explaining principle of parsimony in context).

221. Kenneth A. Bamberger & Deirdre K. Mulligan, *Privacy on the Books and on the Ground*, 63 STAN. L. REV. 247, 258–59 (2011).

V. CONCLUSION

Protecting privacy in an always-on era is very challenging. When individuals are constantly surrounded by devices that might capture their daily routine, conversations, location, imagery, and vital signs, they must have safeguards against misuse of these data. The sectoral approach does little to advance the rationales of protecting privacy in this age, so policymakers must further examine it. And policymakers should strive to embrace other regulatory mechanisms that would better protect sensitive data as sensors become more embedded in our lives. But as illustrated above, technological solutions must also be considered, as they might enhance privacy protection for individuals while preserving the value of data to a greater extent than the current regulatory approaches can. To accomplish this, OSPs might be obliged, or incentivized, to deploy mathematical solutions that will depend on an ex ante evaluation of the probability of data sensitivity.

Data sensitivity will also change with technology. As individuals make more use of IoT technology, including its potential embedment in the public infrastructure, we might divulge more data to both private companies and government agencies than ever before. Thus, any privacy model, including the one proposed in this Article, must be further examined and recalibrated to embed the values that society wishes to protect. For the time being, policymakers must consider requiring OSPs to implement innovative technological and mathematical solutions, such as the proposed framework, to address the profound privacy concerns that emerge from the always-on era.