Preserving Digital Privacy in the Light of Big Data

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Patrick Thomas
Spring, 2021

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date _____
Patrick Thomas

Approved _____ Date _____
Sean Ferguson, Department of Engineering and Society

Abstract

Since entering the Information Age and especially in the past couple of decades, society has become strongly integrated with Internet technologies across a wide variety of fields, such as government, healthcare, and entertainment. In the beginning, the Internet was a network of networks intended to connect internal academic and military networks to support collaborative efforts between institutions. Now, it connects users to online banking, shopping, information, and each other. With the strong potential for commerce on the Internet, business interests, marketing, and advertising have taken over much of the Internet. Towards maximizing the profit potential, these same groups have taken to bulk personal data gathering and analysis without regard for user privacy. Digital privacy has a great deal of importance but has been attacked and diminished on websites and online platforms over recent years. A variety of solutions to help remedy and restore privacy online exist in plenty of forms: government, ad blockers, new communication protocols, pedagogy, and advocacy. Ad blocking and DNS-over-HTTP have a trade-off of helping establish individual privacy at the cost of others. Advocacy and education start at the solution of the problem and are a satisfactory, long-term approach to increasing the value of privacy both to businesses and individuals. Government works to help codify best practices and protect privacy rights at the cost of clarity to businesses. However, there needs to be joint action on multiple fronts that reflects the complex relations better government, Internet business and infrastructure, and users to fully realize online privacy.

Preserving Digital Privacy in the Light of Big Data

The Internet and Internet applications have become intertwined with life in many areas worldwide. Internet-related technologies have explosively grown in the past couple of decades and now connects people to online retailers, banking, social media, and government services. Even though that many of these web applications and online services are free, they often still come with a cost to the end user: personal data. Nowadays, websites and online advertisers may collect large amounts of personal information unbeknownst to the end user to maximize the platform's profit, either by selling the information to marketers and advertisers or scrutinizing the data to tailor their services. This datafication directly harms individual digital privacy and leaves users without control of their online information, and the number of datafication methods is only growing and is fueled by poor online privacy.

Nowadays there is a wide array of methods and tools to gather personal information about an individual user, or at least enough information to uniquely identify that user across websites. The most straightforward method is hidden trackers in ads, which leave cookies or record visits to a website. This provides a way for advertisers to build a history of websites visited by a single user. For example, Eckersley (2010) showed that 94.2% of browsers could be uniquely fingerprinted from just visiting a single URL and that 99.1% of subsequent visits to that same URL could be successfully linked to a previous visit (Eckersley, 2010). Alternatively, a single website like Facebook.com can easily track all activity within its website. Even more advanced ways to gather personal information exist like Carnus, which can identify 97% of installed browser extensions to launch inference attacks (Karami et al., 2020). Beyond web browser privacy is DNS privacy, where Dickinson (2020) argues that "the DNS is one of the most significant leaks of data about an individuals and an organisations activity on the Internet"

(Dickinson, 2020). These methods only getting more advanced, and each of these poses a significant technological challenge to achieving digital privacy.

Digital Privacy and the Commodification of Data

Digital privacy and privacy as a whole have a great deal of importance and motivate the thesis. For this paper, digital privacy, or privacy on the Internet, is defined as an Internet user's right to choose who accesses their data as well and knowledge of who uses their data for what. At the core of this concept is privacy itself. As enumerated by Magi (2011), the benefits range from benefits to the individual (ability to express ideas without self-censorship, affirmation of self-ownership, increased freedom of choice, etc.), personal relationships, and to society as a whole (supports the common good and political activity, acts a layer of protection in government/organization and citizen power imbalances, etc.) (Magi, 2011). McFarland (2012) attributes the importance of privacy to the consequences of a lack thereof and that privacy helps preserve the boundaries of oneself to maintain autonomy (McFarland, 2012). Mai (2016) proposes that there are four forms of privacy: physical privacy, decisional privacy, psychological privacy, and information privacy, which is of particular importance to this paper (Mai, 2016).

Given the importance of privacy, what happened to it, and why do corporations and governments take advantage of personal data? Sarathy and Robertson (2003) highlight two types of privacy problems: *privacy vs. societal good* and *privacy vs. commercial interests* (Sarathy & Robertson, 2003). Users often explicitly trade personal information like names, emails, addresses, and age with websites in exchange for better-personalized services, providing some social good. This exchange often differs greatly from a user's stated views on digital privacy, showing that there is some dissonance between how a user control's their privacy and how a user perceives it (Norberg et al., 2007). Additionally, users are not usually concerned with privacy but

rather with protecting private information about themselves (Waldo et al., 2007). As opposed to explicitly handing over personal information, users also often unknowingly have their web traffic collected and analyzed without consent or notification. Users' personal information is often shared with far more websites than expected, and this knowledge may only be embedded deep in privacy policies. Now large , advances in Big Data have allowed for the transformation of large amounts of data into insights. A problem for businesses and governments with such personal data, regardless of consent, is the moral dilemma of whether they the moral right to leverage such data for commercial interests (Mai, 2016). Another challenge of digital privacy is the complex relations of all of the stakeholders involved. For example, Internet governance in the United States alone is mostly controlled by the multiple levels of government, national and international bodies and committees, advocacy and civil action groups, and a multitude of businesses surrounding the Internet (Internet service providers, infrastructure, etc.) (Carr, 2015). This makes any form of global Internet governance difficult since any single global system ignores the privatized nature of Internet infrastructure and administration (DeNardis & Raymond, 2013).

Restoring Privacy: Solutions to Datafication

This paper will consider several solutions that aim or happen to remedy some aspect of datafication and in turn restore some digital privacy. In particular, DNS-over-HTTPS, ad blockers and browser extensions, pedagogy, government action, and advocacy will each be analyzed in their impact concerning how to support privacy or the fundamental ideals behind privacy.

DNS-over-HTTPS

The Domain Name System (DNS) is a quintessential element of what makes the Internet work as it does today. The DNS is a system of resolving a URL (like virginia.edu) to a computer's IP address (like 128.143.67.11). The original design of DNS called for no encryption at all, and the hierarchical, distributed, public design of it means that it is relatively simple to launch attacks on DNS traffic. This enables mass surveillance due to the widespread usage of DNS (Guha & Francis, 2007). DNS-over-HTTPS (DoH) is a technology that proposes to make the DNS both more secure and confidential (Internet Engineering Task Force, 2020). In essence, DoH aims to encrypt DNS queries, which makes the queries unreadable and immutable from anyone except for the intended recipient. DoH faces a great deal of resistance from national telecommunications and Internet organizations like the Cellular Telecommunications and Internet Association (CTIA), the Internet & Television Association (NCTA), and the United States Telecom Association (US Telecom) because DoH disables marketing from regular DNS queries and of fears that a single large company like Google would become the only DoH provider, giving them a monopoly on the information (CTIA et al., 2019; Lee, 2019). DoH prevents third-party snooping of DNS queries; however, the top-level DNS resolver can still read and market data from queries. This means that whoever is the first to widely support and enable DoH is the sole marketer of such data, further stoking fears that Google is trying to gain sole control of DNS query information. Likewise, national broadband and Internet providers are unlikely to support this through infrastructure upgrades simply since it doesn't benefit them. DoH in its current form has to be manually enabled and is heavily reliant on both operating system and browser support, making this solution only available to those who both are technically capable of enabling it and have operating system support.

Ad Blockers and Browser Extensions

A well-known and widely-used method of preserving privacy is through browser extensions and especially ad blocking browser extensions. A 2017 study estimated that there are 500 million ad block users worldwide, and this number is expected to grow by 40% every year (Garimella et al., 2017). Ad blockers directly help privacy for an individual by stopping online ads and trackers from functioning, making the user's online traffic much harder to capture, which helps to affirm the self-ownership of personal data as well as promote autonomy online.

While ad blocking may be a good short-term solution to get some privacy, their usage fuels an ongoing ad blocking arms race. According to Nithyanand et al. (2016), an estimated 6.7% of sites in the Alexa Top-5K website ranking employ some sort of counter-ad blocking technology (Nithyanand et al., 2016). Despite ad blockers being perhaps the easiest program to install because of their close integration with modern web browsers, some barriers to usage are their discoverability and knowing where to find such extensions. Additionally, ad blocker adoption is mostly motivated by user ad annoyance and much less so from privacy or security-related desires (Hessen et al., 2020). Ad blocker usage can even counter-intuitively give trackers more information towards uniquely identifying an individual in certain cases (Eckersley, 2010). Finally, using privacy-enhancing browser addons helps achieve personal digital privacy at the cost of others, as now there is a smaller pool of candidates for others to be identified from (Eckersley, 2010).

Government

Laws and regulations are very important in limiting what businesses do with personal information, thus giving government a large role in controlling datafication. A largely successful and recent such regulation is the General Data Protection Regulation (GDPR) passed by the

European Union, which requires rights such as consent, the right to be forgotten, and the right to access for websites that interact with European Union users. Recent similar work to the GDPR is from Virginia and was just passed into law on March 2, 2021: the Consumer Data Protection Act (Marsden, 2021). Legal frameworks like the GDPR and the California Consumer Protection Act (CCPA) are strong in that they directly impose limits on businesses and encode peoples' rights to their data and data choices, among other actions.

These regulations are not without their issues. Full GDPR compliance comes with a host of implementation-specific issues; for example, full compliance with the right to be forgotten means that a user's data must also be removed from data backups, a non-trivial and computationally expensive task. GDPR also has problems with defining what personal information is required for business operations; an example is an online retailer that collects asks for its users' birthdays. On its own, collecting birthdays for no business reason is not justifiable under GDPR. However, if the retailer were to justify it by sending out birthday coupons, they can then legally justify collecting that personal information. CCPA has its own pitfalls and suffers from poor clarity; businesses have trouble knowing what certain parts of CCPA apply to their business and whether some practice is CCPA-compliant or not. Finally, both the GDPR and CCPA are also meant to be as nonspecific as possible regarding implementations and technology. This is to account for the rapid pace at which computer technology is developed, but it also means that their requirements are intentionally ambiguous and do not directly map to concrete software requirements.

Despite some of the pitfalls of GDPR and CCPA, they are relatively well-known and are successful in raising concern over privacy-related issues on the Internet and greatly advance online privacy rights, advancing citizen political engagement in privacy issues. As mentioned

previously, GDPR and CCPA also are effective in enumerating what specific rights citizens have over their personal data online, directly supporting the fundamentals of privacy.

In a more broad sense, government is not without its issues. The United States, for example, recognizes that privacy ultimately hurts large Internet businesses the most and cuts into profits from advertising and marketing. Thus, Governments also have something to gain from poor privacy regulation too, and organizations like the National Security Agency have vested interests in mass surveillance.

Advocacy

Advocacy and civil action groups serve an important role in the ecosystem of Internet administration and infrastructure, lawmakers, engineers, and more. Groups like the Center for Democracy and Technology help push and guide legislation that supports digital privacy rights while fighting against laws that roll back or reduce online protections. Advocacy helps focus public opinion on privacy issues and is some of the most important actors in shaping public opinion of privacy (Waldo et al., 2007). Advocacy groups often work with columnists or journalists in spreading the word about privacy issues towards motivating the public. Advocacy has its limits: limited funding for advocacy groups means that corporations are easily able to fund opinion surveys to push their agendas (Waldo et al., 2007). In all, rather than try and stop the flow of personal information like ad blockers and DoH, advocacy affects the course of corporations, businesses, and law in direct support of privacy as a whole.

Pedagogy

A different way to approach datafication is through education. Ideally, we will find methods that solve the core of the problem of datafication rather than handle the consequences thus so far, and collegiate-level, privacy-focused education helps to solidify those concepts early.

For undergraduate computer science students, adding material like Privacy by Design to the required curriculum would help expose students to the values of privacy as well as the implications, trade-offs, and goals when building a private system (Langheinrich, 2001). There are a set of challenges associated with ethics and ethics-based topics like privacy, as ethics topics often are hard to convey or are misinterpreted by students (Holsapple et al., 2012). Case studies are common but often just illustrate negative examples or what not to do (Shilton et al., 2020). Shilton et al. (2020) developed a board game to not only walk through Privacy by Design but also integrate other elements of software engineering, allowing students to safely engage with and apply privacy concepts to a simulated project (Shilton et al., 2020). Shilton's work also demonstrably helped students approach ethics in computing problems and more importantly increased student interest in ethics issues. Similar to advocacy, changes in educations towards privacy also affect privacy on a higher level, informing better design for the future.

Discussion

All of the aforementioned solutions have their merits and pitfalls. DNS-over-HTTP and browser extensions exist within the domain of computer technology and are self-serving but also render immediate, personal benefits to the user. Government and law-making is a slow process but yields a generally uniform set of rules to guide later design decisions. Advocacy is great in shaping legislation and getting people active in privacy. Finally, education is a great long-term solution, but its benefits can be hard to achieve. Advocacy and education are perhaps slower processes than government but strike directly at the core of privacy by affecting legislation and design.

Ad blocking, when viewed as a privacy-enhancing technology rather than an annoyance-blocking one, has several downsides. Ad blocker usage is preceded by an operating system and

browser support, limiting the number of people who can access the technology. Additionally, ad blocker usage can make others more easily identifiable and well as paradoxically make the ad block user more easily identifiable. Furthermore, ad blockers have a profound negative effect on websites that rely on advertising revenue and have limited other options for monetization, like online journalism. Given the downsides, ad blocking is not a clean solution that helps us immediately gain digital privacy, and it perhaps never aimed to be as such; an ad blocker's main goal, after all, is to stop ads. Arguably, websites have an ethical and moral obligation to monetize in ways that further societal good and therefore support digital privacy, and advertising based on personal information without express consent is antithetical to this. This introduces the idea that we should move beyond traditional advertising, but uprooting and changing the digital advertising industry would require a great deal of pressure from multiple groups.

DNS privacy, like DoH, differs from ad blocking in its motivation, better privacy for otherwise entirely unsecured DNS queries. DNS privacy, however, achieves a similar effect on the end user in that it is a relatively small fix for a much larger issue. Furthermore, the existing DNS network lacks the infrastructure and upgrades necessary to handle DoH save for tech giants who are first to implement and support DoH, like Google.

Both ad blockers and DNS privacy help maintain more control over personal information. Despite the downsides of either technology, control over personal information is important since it is critical to supporting digital privacy. These technologies also forcefully opt out of tracking and datafication in situations where it may be impossible, giving its users a greater degree of choice in how their data is processed. A solution that changes the course of datafication rather than cleans up its consequences would ultimately serve Internet users better.

Government emerges as a good mediator of digital privacy between citizens and businesses. Regulations like the GDPR and CCPA are good at enforcing ideals of the growing sentiment against datafication and help to codify privacy requirements in software engineering applications. While this role was previously fulfilled by professional computing societies and enterprise-level policy, the government has the power to make rules that are both ubiquitous and enforceable. These frameworks and laws codify the previously mentioned privacy tenets. Additionally, these frameworks help bring privacy issues to the forefront for many, increasing public perception of privacy and hopefully increasing citizen engagement in online privacy issues.

Pedagogical methods and advocacy both help to stem the source of the datafication problem that is pervasive on the Internet. Although it is slow, educating engineering and business students about the greater ethical problems in the systems that they design, build, and use is the best long-term solution to help digital privacy. These methods are the best among the aforementioned that stir political engagement among Internet users and those who build Internet systems. That being said, it still has to happen and changes in curriculum need to be made, and changes would not be apparent for years. Advocacy helps to shape other societal factors like government, thus these two solutions together approach the Internet through the complex stakeholder system of government and business that it exists in. However, none of the digital privacy solutions can become the norm without widespread support, and the solution can not precede the perception of a problem. Advocacy and education help to especially change this as it helps to bring the core issue of datafication and how it attacks privacy to the general public. Advocacy group collaboration with columnists and journalists also helps achieve this goal and exposing the current state of the system. Education helps to achieve this from the inside of

software engineering, empowering engineers with the toolset they need to navigate not only digital privacy but other ethical issues. While education and advocacy do not directly or immediately affect digital privacy, they do entirely support the ideals of digital privacy, like affirming self-ownership, supporting freedom of choice, and encouraging political engagement.

Conclusion

Overall, a combination of education and advocacy are the most promising solutions moving forward, and advocacy also includes government action through guiding legislation. These methods directly support digital privacy and its fundamental ideals. They are not mere patches on the resulting problems like ad blockers and communication protocols are, and they help to shape society rather than clean up the mistakes. Unfortunately, there is limited literature as to how effective such changes would be in the long run, so this research is by no means comprehensive. While ad blockers and DoH face challenges and have their downsides, it is also hard to discount those solutions completely since users can directly affect websites that abuse poor online privacy and datafication. While beyond the scope of this paper, moving to an alternative model for online advertising may still be very promising.

References

Carr, M. (2015). Power Plays in Global Internet Governance. *Millennium*, *43*(2), 640–659. https://doi.org/10.1177/0305829814562655

CTIA, NCTA, & US Telecom. (2019, September 19). *DNS-over-HTTP Letter to Congress*. https://www.ncta.com/sites/default/files/2019-09/Final DOH LETTER 9-19-19.pdf

DeNardis, L., & Raymond, M. (2013). *Thinking Clearly About Multistakeholder Internet Governance* (SSRN Scholarly Paper ID 2354377). Social Science Research Network. https://doi.org/10.2139/ssrn.2354377

Dickinson, S. (2020, December 14). *DNS Privacy Project—The Problem*. DNS Privacy Project. https://dnsprivacy.org/wiki/display/DP/DNS+Privacy+-+The+Problem

Eckersley, P. (2010). How Unique Is Your Web Browser? In M. J. Atallah & N. J. Hopper (Eds.), *Privacy Enhancing Technologies* (pp. 1–18). Springer. https://doi.org/10.1007/978-3-642-14527-8_1

Garimella, K., Kostakis, O., & Mathioudakis, M. (2017). Ad-blocking: A Study on Performance, Privacy and Counter-measures. *Proceedings of the 2017 ACM on Web Science Conference*, 259–262. https://doi.org/10.1145/3091478.3091514

Guha, S., & Francis, P. (2007). Identity Trail: Covert Surveillance Using DNS. In N. Borisov & P. Golle (Eds.), *Privacy Enhancing Technologies* (pp. 153–166). Springer. https://doi.org/10.1007/978-3-540-75551-7_10

Hessen, M., Hyll, M. C., Momeninasab, L., Padala, A., & Ulvund, L. (2020). *Motivations for ad-block usage among digital newspaper readers*. 4.

Holsapple, M. A., Carpenter, D. D., Sutkus, J. A., Finelli, C. J., & Harding, T. S. (2012). Framing Faculty and Student Discrepancies in Engineering Ethics Education Delivery. *Journal*

of Engineering Education, *101*(2), 169–186. https://doi.org/10.1002/j.2168-9830.2012.tb00047.x

Internet Engineering Task Force. (2020, March 17). *DNS Over HTTPS (doh)*. IETF Datatracker. https://datatracker.ietf.org/wg/doh/about/

Karami, S., Ilia, P., Solomos, K., & Polakis, J. (2020). Carnus: Exploring the Privacy Threats of Browser Extension Fingerprinting. *Proceedings 2020 Network and Distributed System Security Symposium*. Network and Distributed System Security Symposium, San Diego, CA. https://doi.org/10.14722/ndss.2020.24383

Langheinrich, M. (2001). Privacy by Design—Principles of Privacy-Aware Ubiquitous Systems. In G. D. Abowd, B. Brumitt, & S. Shafer (Eds.), *Ubicomp 2001: Ubiquitous Computing* (Vol. 2201, pp. 273–291). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45427-6_23

Lee, T. B. (2019, September 30). *Why big ISPs aren't happy about Google's plans for encrypted DNS*. Ars Technica. https://arstechnica.com/tech-policy/2019/09/isps-worry-a-new-chrome-feature-will-stop-them-from-spying-on-you/

Magi, T. J. (2011). Fourteen Reasons Privacy Matters: A Multidisciplinary Review of Scholarly Literature. *The Library Quarterly*, *81*(2), 187–209. https://doi.org/10.1086/658870

Mai, J.-E. (2016). Big data privacy: The datafication of personal information. *The Information Society*, *32*(3), 192–199. https://doi.org/10.1080/01972243.2016.1153010

Marsden, D. W. (2021, February 8). *SB 1392 Consumer Data Protection Act; establishes a framework for controlling and processing personal data.* Virginia's Legislative Information System. https://lis.virginia.gov/cgi-bin/legp604.exe?211+sum+SB1392

McFarland, M. (2012, June 1). *Why We Care about Privacy*. Why We Care about Privacy. https://www.scu.edu/ethics/focus-areas/internet-ethics/resources/why-we-care-about-privacy/

Nithyanand, R., Khattak, S., Javed, M., Vallina-Rodriguez, N., Falahrastegar, M., & Powles, J. E. (2016). Adblocking and Counter-Blocking: A Slice of the Arms Race. *University of Cambridge*, 7.

Norberg, P. A., Horne, D. R., & Horne, D. A. (2007). The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors. *Journal of Consumer Affairs*, *41*(1), 100–126. https://doi.org/10.1111/j.1745-6606.2006.00070.x

Sarathy, R., & Robertson, C. J. (2003). Strategic and Ethical Considerations in Managing Digital Privacy. *Journal of Business Ethics*, *46*(2), 111–126. https://doi.org/10.1023/A:1025001627419

Shilton, K., Heidenblad, D., Porter, A., Winter, S., & Kendig, M. (2020). Role-Playing Computer Ethics: Designing and Evaluating the Privacy by Design (PbD) Simulation. *Science and Engineering Ethics*. https://doi.org/10.1007/s11948-020-00250-0

Waldo, J., Lin, H. S., & Millett, L. I. (Eds.). (2007). *Engaging Privacy and Information Technology in a Digital Age* (p. 496). National Academies Press. https://doi.org/10.17226/11896