TRACK THE PLANET: A WEB-SCALE ANALYSIS OF HOW ONLINE BEHAVIORAL ADVERTISING VIOLATES SOCIAL NORMS

Timothy Libert

A DISSERTATION

in

Communication

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2018

Supervisor of Dissertation

_____

Dr. Victor Pickard, Associate Professor of Communication

Graduate Group Chairperson

_____

Dr. Joseph Turow, Robert Lewis Shayon Professor of Communication

Dissertation Committee
Dr. Michael X. Delli Carpini, Walter H. Annenberg Dean of The Annenberg School for Communication
Dr. Joseph Turow, Robert Lewis Shayon Professor of Communication
Dr. Jonathan M. Smith, Olga and Alberico Pompa Professor of Engineering and Applied Science
Dr. Guobin Yang, Professor of Communication and Sociology

Acknowledgments

Thanking everybody who has helped me reach this milestone is an impossible task and I will undoubtedly leave out people who deserve recognition. Nonetheless, the following is my attempt at doing so and I apologize in advance for the many omissions I have made.

First, I owe a debt to the entire staff at the Annenberg School who have been helpful and gracious over the entirely of my graduate school career. I especially need to thank everybody in the IT department for supporting my research activities even when they generated copious volumes of security alerts from various corners of Penn's network defenses. As a former university IT administrator myself, I would have *never* had the patience to support a student doing what I did and would have terminated network access without a second thought. It is my good fortune that the Annenberg IT staff are more far more understanding and generous than I would have been if our roles were reversed.

I extend a broad thanks to my fellow students who challenged me to defend my positions in the classroom and provided me with a sounding board for new ideas

outside of it. However, beyond intellectual support, is was the emotional support of my colleagues which helped get me through periods of self-doubt and frustration.

I thank all of the faculty I studied with during my time at Annenberg. The diverse range of methods and theories I was exposed to in the classroom made me a better scholar and gave me greater insights into, and respect for, the work of others. Beyond the substance of the courses I took, being exposed to many unique styles of pedagogy has influenced my own approach.

I thank Prof. Sandra González-Bailón for hosting the DIMENET meetings, where I got to hear excellent feedback from my peers as well as learn more about a range of exciting methods. Most important to me, Prof. González-Bailón fostered an environment which was always brutally honest, but never mean-spirited. Prof. González-Bailón also gave me perhaps the most useful single comment on a paper I received while at Annenberg which was to "cut out all the bullshit".

In addition to being an excellent committee member, Prof. Guobin Yang helped me grow as a lecturer and showed me how to keep a classroom of undergraduates engaged for an entire hour, week after week. Prof. Yang taught me an enormous amount about Chinese media and Internet regulation, but more importantly he taught me about the unique cultural factors which give "Chinese Characteristics" to a range of social and regulatory issues. In the coming decades I expect that Prof. Yang's influence will shape my understanding of world events.

I thank Dean Michael X. Delli Carpini for providing valuable feedback as a

committee member and for being an excellent dean during the entirely of my time at Annenberg. Dean Delli Carpini has always demonstrated a sincere interest in what students are working on, has been genuinely curious about our work, and has been deeply invested in our success. Likewise, the overall 'vibe' of an institution always draws in some way from the personality at the top and Dean Delli Carpini's relaxed attitude and good nature has had a positive impact on the entire school.

Prof. Monroe Price has provided me with countless opportunities to engage with non-academic stakeholders, work on human rights issues, and use my scholarship as a means to assist with civil society campaigns. At times when I wondered what the real-world impact of my scholarship was, an email from Prof. Price introducing me to an activist would somehow land in my inbox, reminding me of why I was doing the work I was doing. Likewise, by introducing me to his global network of collaborators, Prof. Price helped me develop the type of cosmopolitan perspective needed to succeed in a connected world.

Among the many people Prof. Price introduced me to, the Ranking Digital Rights (RDR) team and the leadership of the Center for Media, Data, and Society at the Central European University (CEU) deserve particular mention. Working with Rebecca MacKinnon from the earliest days of RDR was a formative experience which gave me deep insights into activism, policy, and the ways in which the Internet is a truly global medium. Likewise, spending years reading privacy policies for RDR inspired me to figure out how to get computers to do the task for me and contributed

greatly to this dissertation. RDR team member Allon Bar was not only an excellent colleague, but became an even better friend to myself and Dana when we arrived in Berlin. At CEU in Budapest, Prof. Kate Coyer and Éva Bognár took me on an exciting journey to deploy wifi to refugees and brought me to the most memorable Thanksgiving party of my life. I have never been so proud to be a part of any organization as when I learned the Hungarian government sought to silence CEU. Having spent time working with Kate and Éva I know the effort is doomed to failure.

Prof. Joe Turow was an invaluable member of my committee, challenging me to justify how my methodological choices related to the larger theories with which I was engaging. Prof. Turow prevented me from getting lost in methods at the expense of seeing the big picture. Likewise, Prof. Turow has provided a source of privacy scholarship which has influenced my own thinking greatly, and serves as an example of how scholars can have an important role in shaping public understanding of privacy issues.

Assistant Dean for Graduate Studies Joanne Murray deserves thanks for too many things to count, among them being able to absorb an unending stream of complaints with good humor and inhuman patience, making sure that from the time I walked in the door I was focused on walking out of the door, knowing every single secret of every regulation at Penn so I didn't need to, answering emails with dumb questions I should have known the answer to, and most importantly, being the first person at Penn to give me a chance by putting my application in front of

the admissions committee. Joanne is the real MVP.

I thank Prof. Jonathan Smith of the engineering department for treating me as one of his own, making time to have meetings with me, inviting me into his classroom, and serving as a methodology supervisor for my research. Most importantly, Prof. Smith gave me confidence that my work was making a contribution to the field of computing as well as communication, and that I should be comfortable presenting my findings to both groups. Although imposter syndrome is endemic in the academic pursuit, Prof. Smith has allowed me to consider the possibility that I may, in fact, be a legitimate researcher, and for that I am thankful.

On a practical level, Prof. Smith directly facilitated the production of this dissertation by giving me access to significant computing resources which I attempted to (ab)use to their full potential. Without the resources provided by Prof. Smith the data for this dissertation would never have been collected or processed in a remotely reasonable time period, nor would many aspects of this dissertation been financially feasible.

I thank Prof. Victor Pickard for being an excellent advisor, buying me countless meals, being available 24x7 for answering questions ranging from the petty to the substantive, helping make sure minor problems didn't end up being major ones, making sure that manuscripts which started as angry political diatribes turned into thoughtful scholarship, and showing me how scholarship can support activism. Most of all, I thank Prof. Pickard for placing enormous trust in me to set my own

goals and timelines and follow through on them. Knowing that Prof. Pickard was trusting me to stay on top of my projects made me work harder to make sure his trust was not misplaced.

While the people mentioned so far have helped make me a better scholar, it is my family who have made me a better *person*. My mother-in-law Barbara never hesitated to provide professional and personal guidance and made sure that despite being a poor grad student I continued to enjoy the weekly brunches and cocktails that are the foundation of civilized life. My mother Anne instilled in me an overabundance of confidence, stubbornness, and dedication to fighting on behalf of others. It is those traits which are the bedrock of this dissertation and the underlying motivation for my work. On a practical level, my mother read every word of every paper I've ever written (including this dissertation), and has provided over thirty years of professional copy-editing services completely gratis.

Most importantly, I thank my wife Dana for her unending support, patience, and personal sacrifice. If I have ever come across as smart or insightful it is because I first shared my thoughts with Dana and she graciously filtered out all of the bad ideas, helped refine the good, and contributed her own insights. In particular, the chapter on medical privacy in this dissertation was originally her idea. Despite having a comfortable life in New York, Dana joined me on a long journey from Philadelphia, to Budapest, to Berlin, back to Philadelphia, and to the eternally grey and rainy Oxford. The journey has been long and fascinating, but disruptive

and exhausting as well. Along the way I often wanted to quit, but Dana continued to support and encourage me despite the fact that me giving up would have brought more sanity and stability to her own life.

Last, I thank my son Jacob for napping long enough for me to finish writing this damn thing and reintroducing me to my love of music. Hopefully by the time he reads this the state of online privacy will have gotten marginally better and he won't be fighting off Terminators, rising seas, and even more unimaginably invasive forms of advertising.

ABSTRACT

TRACK THE PLANET: A WEB-SCALE ANALYSIS OF HOW ONLINE BEHAVIORAL ADVERTISING VIOLATES SOCIAL NORMS

Timothy Libert

Dr. Victor Pickard

Various forms of media have long been supported by advertising as part of a broader social agreement in which the public gains access to monetarily free or subsidized content in exchange for paying attention to advertising. In print- and broadcast-oriented media distribution systems, advertisers relied on broad audience demographics of various publications and programs in order to target their offers to the appropriate groups of people. The shift to distributing media on the World Wide Web has vastly altered the underlying dynamic by which advertisements are targeted. Rather than rely on imprecise demographics, the online behavioral advertising (OBA) industry has developed a system by which individuals' web browsing histories are covertly surveilled in order that their product preferences may be deduced from their online behavior. Due to a failure of regulation, Internet users have virtually no means to control such surveillance, and it contravenes a host of well-established social norms.

This dissertation explores the ways in which the recent emergence of OBA has come into conflict with these societal norms. Rather than a mere process for tar-

geting messages, OBA represents a profound shift in the underlying balance of power within society. This power balance is embedded in an information asymmetry which gives corporations and governments significantly more knowledge of, and power over, citizens than vice-versa. Companies do not provide the public with an accounting of their techniques or the scale at which they operate.

In order to shed light on corporate behavior in the OBA sector, two new tools were developed for this dissertation: `webXray` and `policyXray`. `webXray` is the most powerful tool available for attributing the flow of user data on websites to the companies which receive and process it. `policyXray` is the first, and currently only, tool capable of auditing website privacy policies in order to evaluate disclosure of data transfers to specific parties. Both tools are highly resource efficient, allowing them to analyze millions of data flows and operate at a scale which is normally reserved for the companies collecting data. In short, these tools rectify the existing information asymmetry between the OBA industry and the public by leveraging the tools of mass surveillance for socially-beneficial ends.

The research presented herein allows many specific existing social-normative concerns to be explored using empirical data in a way which was not previously possible. The impact of OBA on three main areas is investigated: regulatory norms, medical privacy norms, and norms related to the utility of the press. Through an examination of data flows on one million websites, and policies on 200,000 more, it is found in the area of regulatory norms that well-established Fair Information Prac-

tice Principles are severely undermined by the self-regulatory "notice and choice" paradigm. In the area of informational norms related to personal health, an analysis of data flows on 80,000 pages related to 2,000 medical conditions reveals that user health concerns are shared with a number of commercial parties, virtually no policies exist to restrict or regulate the practice, and users are at risk of embarrassment and discrimination. Finally, an analysis of 250,000 pages drawn from 5,000 U.S.-based media outlets demonstrates that core values of an independent and trustworthy press are undermined by commercial surveillance and centralized revenue systems. This surveillance may also transfer data to government entities, potentially resulting in chilling effects which compromise the ability of the press to serve as a check on power.

The findings of this dissertation make it clear that current approaches to regulating OBA based on "notice and choice" have failed. The underlying "choice" of OBA is to sacrifice core social values in favor of increased profitability for primarily U.S.-based advertising firms. Therefore, new regulatory approaches based on mass surveillance of corporate, rather than user, behaviors must be pursued. Only by resolving the information asymmetry between the public, private corporations, and the state may social norms be respected in the online environment.

Contents

Chapter 1

Introduction

> "A wonderful fact to reflect upon, that every human creature is constituted to be that profound secret and mystery to every other. A solemn consideration, when I enter a great city by night, that every one of those darkly clustered houses encloses its own secret; that every room in every one of them encloses its own secret; that every beating heart in the hundreds of thousands of breasts there, is, in some of its imaginings, a secret to the heart nearest it!" (Dickens, 1859)

In the passage above, a traveller in Charles Dickens' 1859 classic *A Tale of Two Cities* reflects on the "wonderful fact" that all people are in some way an enigma unto themselves, unknown even to those closest to them. It is impossible for the traveller to peer into the "darkly clustered houses", and the lives of the inhabitants are forever a mystery to him. However, the "beating hearts of hundreds of thousands

of breasts" are now laid bare before the billions of websites which permeate all sectors of society and the economy. These websites harbor a massive surveillance infrastructure which has denuded the landscape of secrets and undermined a wide range of privacy norms which predate Dicken's novel by millennia.

When a bedroom door is closed today, the occupant may be physically alone, but she or he is often connected to vast swaths of humanity via the World Wide Web. In its short history, the web has rapidly evolved from an interactive hypertext media into a container media which has become the primary conduit for the delivery of not only text, but images, music, television, film, games, as well as augmented and virtual realities. While the rise of mobile devices is sometimes viewed as a turn away from the web, in reality the most popular applications contain web browser functionality and are often built on web infrastructure. The web has become so ubiquitous that any distinctions between the "offline" and "online" worlds have become essentially meaningless as nearly all media and communication is subsumed into Hypertext Transfer Protocol (HTTP) traffic.

As traditional forms of media have moved to the web, they have adapted to new technological and economic dynamics, the biggest of which is a fundamental shift in the relationship between publishers and media consumers. Whereas broadcast and print media operated primarily as a means for one-way communication, the web offers a means for audiences to communicate with each other and with media providers. Some of these communications are visible, as when a user shares a social

media post with friends or posts a comment on a newspaper article. However, the vast majority of user communications take the form of low-level network traffic which is typically hidden from view. Such traffic reveals the IP addresses, browser information, and cookies of a user visiting a website, allowing them to be identified and tracked. Whereas the operator of a radio or television station has no means of verifying who, if anybody, is listening or watching, the owner of a website may know *exactly* who is visiting.

Traditional media have depended on advertising revenue for much of their history. In the one-to-many communication model, advertisements are tailored for specific audiences only to the degree that certain publications or programs are know to be popular with certain groups, such as young men, sports fans, or retirees. The best means of measuring the potential impact of such advertisements is sales volume and brand awareness. While there are a host of substantive objections to the commercialization of media, there are few, if any, major privacy objections. Dicken's vision of rooms enclosing secrets is not radically changed if the rooms contain traditional television sets.

In contrast, because the web facilitates monitoring the actions of specific users, advertisers are no longer forced to pitch the merits of their wares to roughly-clustered, faceless masses. Rather, by leveraging the hidden communications contained in network request traffic, they are able to isolate specific users. They are able to monitor these users as they traverse the web, developing deep and rich pro-

files of the user's interests and behaviors. Leveraging this data they are able to "target" users for "tailored" advertisements and determine if they have interacted with the advertisement and made a purchase. Due to increased network speeds, the proliferation of mobile devices, and plummeting costs of data storage, both the volume of information ingested and the time it is retained are virtually limitless. This new approach, called online behavioral advertising (OBA), is a fundamental departure from all that came before it.

OBA has become such a dominant force on the web that wherever the web goes, for-profit surveillance follows. Spheres of life which were previously far from the commercial realm are now subject to an omnipresent surveillance apparatus designed to monitor visits to one website in order to target an advertisement on another. Social contexts with well-developed informational norms, such as those related to medical issues, have been grossly violated by OBA. Users may be discriminated against in the marketplace based on their web browsing histories. Most troublesome, government spy agencies have been able to repurpose commercial surveillance tools in order to conduct mass warrantless surveillance. Thus, while advertisements are frequently viewed as tolerable annoyances, the techniques by which they are delivered today ultimately have a fundamental impact on relationships among citizens, corporations, and the state.

This dissertation explores the ways in which the recent emergence of OBA has come into conflict with long-established and closely held societal norms. Rather

than a mere process for "tailoring" messages, OBA represents a profound shift in the underlying balance of power within society. This power balance is embedded in an information asymmetry which gives corporations and governments significantly more knowledge of, and power over, citizens than vice-versa. Companies do not provide the public with an accounting of their techniques or the scale at which they operate.

In order to shed light on corporate behavior in the OBA sector, two new tools were developed for this dissertation: `webXray` and `policyXray`. `webXray` is the most powerful tool available for attributing the flow of user data on websites to the companies which receive and process it. `policyXray` is the first, and currently only, tool capable of auditing website privacy policies in order to evaluate disclosure of data transfers to specific parties. Both tools are highly resource efficient, allowing them to analyze millions of data flows and operate at a scale which is normally reserved for the companies collecting data. In short, these tools rectify the existing information asymmetry between the OBA industry and the public by leveraging the tools of mass surveillance for socially-beneficial ends.

In the same way that inventing a telescope allows one to ask new questions of the heavens, the development of these tools has facilitated the pursuit of understanding the social impact of OBA in a way which would otherwise be impossible. The research presented herein allows many specific existing social-normative concerns to be explored using empirical data in a way which was not previously possible.

The impact of OBA on three main areas is investigated: regulatory norms, medical privacy norms, and norms related to the utility of the press. These investigations reveal that the problem of OBA is not mere invasive advertising, it is the erosion of well-founded social norms.

The findings of this dissertation make it clear that current approaches to regulating OBA based on "notice and choice" have failed. Given the huge volume of surveillance mechanisms on the web today it is technically infeasible for users to be properly notified. Likewise, the underlying "choice" of OBA is to sacrifice core social values in favor of increased profitability for primarily U.S.-based advertising firms. Therefore, new regulatory approaches grounded in social norms must be pursued. These approaches must be enforced using technological tools which monitor the behavior of companies in the OBA space as rigorously and relentless as such companies monitor users. It is time to turn the power of OBA against itself.

This chapter will first provide further background on OBA. Second, the extant state of research will be briefly discussed with attention paid to current gaps in understanding and methodology. Third, the technical underpinnings of OBA are detailed in order that the significant methodological advancements used for this dissertation may be contextualized. Finally, the chapter concludes with a preview of the findings of several case studies.

Online Behavioral Advertising

For many people the cost of accessing the web is largely limited to purchasing a computing device and paying for network access. Nearly unlimited volumes of webpages are freely available without registration or payment due to the fact that otherwise free content is largely subsidized by advertising. As noted above, the provision of content in exchange for viewing advertisements is nothing new: the practice is well-established in mass-media such as newspapers, magazines, radio, and television. The key difference in the online space is that advertisements are often not directed at mass audiences. Rather, advertisements are tailored to the interests of specific *individuals* based on their behavior. This model is known as online behavioral advertising (OBA) and represents a profoundly radical departure from historically accepted commercial practices.

OBA is premised on the fact that showing users advertisements related to traits inferred from online behavior (e.g. websites visited, topics "liked" on social media, etc.) increases the likelihood a given user will choose to follow a web link to a specific advertisement. For example, a user looking up information on infant nutrition may be shown an advertisement for baby food, and a dog owner who "likes" a veterinarian may see advertisements for flea collars. The goal of delivering advertisements targeted in this way is rooted both in satisfying the needs of advertisers and optimizing the use of advertising space by content providers.

The act of clicking on an advertisement is the core action by which money

changes hands in the OBA paradigm. In this "pay-per-click" (PPC) model, advertisers only pay when a user engages directly with their advertisements.[1] On the side of advertisers, a click means a user is choosing to learn more about a product, thus generating a sales lead - this is often the *only* time an advertisers will pay for the advertisement. In theory, this action is measurable and traceable in a way that mass-media advertising is not, demonstrating the value of "ad-spend" to a client. On the publisher side, there is a limited amount of space for advertisements on each page, using an OBA approach, the advertisements on any given page are those most suited for the specific visitor - thereby ensuring that a limited resource is used in the most efficient way possible.

While OBA provides clear commercial benefits, such benefits come at a significant cost to personal privacy. In order to continually refine the effectiveness of ad targeting, vast volumes of data are ingested by OBA companies so that they can fine-tune their behavioral targeting. Large OBA companies such as Google, Facebook, and Twitter process information on user interests gleaned from search terms, private messages, and social media posts. However, there are many online advertising companies who lack large product portfolios from which to gain insights into user behavior. These companies often rely on monitoring the websites users are visiting in order to gain behavioral insights, a technique *also* used by the major companies mentioned above.

[1]There is also a "pay by mille" (PPM) model by which advertisers pay for the number of people who see an advertisement regardless of clicks, but it is less popular online.

Monitoring the pages users visit is often accomplished by inserting computer code into web pages which allows users to be "tracked" as they traverse the web. This process is called "third-party web tracking" and is reliant on a number of features of web pages which were developed for increased functionality, but have since been re-appropriated in service of the surveillance of user habits. Publishers may be paid to place such code on websites, use it for free audience metrics (e.g. Google Analytics), or to add social media sharing functionality. Regardless of the motivation, when users load a given website they are often interacting with many unseen parties.

The broader societal impacts of OBA are not yet understood and there is an extensive academic literature which address issues of privacy broadly, and on the web specifically. Due to a variety of unsettled policy questions regarding the practices involved in OBA, the literature is often directly engaged with public debates.

Extant Literature

There are two main schools of research on privacy and online behavioral advertising (OBA). The first school draws primarily from the social sciences and is rooted in normative evaluations of the legal and ethical impacts of OBA along with documentation of how the practice is viewed by the public. The second school is rooted in computer science and is most focused on documenting the evolving mechanisms employed in OBA and how widespread such mechanisms are. While there is ob-

vious thematic overlap between the two schools, methodological boundaries have prevented deeper integration between social-normative and technical approaches.

Within the social sciences, there are a number of influential theories regarding the role of privacy in normatively-defined contexts (Nissenbaum, 2004; Turow and Hennessy, 2007; Nissenbaum, 2010), democratic contexts (Westin, 1967, 2003), and legal contexts (Solove, 2006, 2012). Ample survey research has shown privacy concerns are not limited to social theorists: it is now well established that online privacy represents an area of significant and sustained concern for the public (Ackerman et al., 1999; Turow, 2003; Acquisti and Gross, 2006; Turow et al., 2009; Hoofnagle et al., 2012; Madden et al., 2012; Fox and Duggan, 2013; Turow et al., 2014; Madden, 2014). Furthermore, there is a large body of social science literature which has documented the endemic failure of so-called "privacy policies" to inform the public of the extent and nature of data collection techniques used in online behavioral advertising (Graber et al., 2002; Blanke, 2006; Milne et al., 2006; McDonald and Cranor, 2008; Barocas and Nissenbaum, 2009; McDonald et al., 2009; Solove, 2012; Barocas and Nissenbaum, 2014; Reidenberg et al., 2014). Overall, this literature identifies normative issues, documents these issues are recognized by the public, and reveals how extant policy and governance mechanisms fail to meet normative goals.

Within the computer science literature, focus has been given to the technical means by which user data is transferred to third parties and sustained research has

documented the programming techniques by which users' computers may be manipulated into facilitating "tracking" within and between browsing sessions (Felten and Schneider, 2000; Jackson et al., 2006; Eckersley, 2010; Jang et al., 2010; Yen et al., 2012; Acar et al., 2013). A related strand of research has conducted censuses to determine how widespread such techniques are across the web (Krishnamurthy and Wills, 2006, 2009; Krishnamurthy et al., 2011; Roesner et al., 2012; Mayer and Mitchell, 2012; Castelluccia et al., 2013; Acar et al., 2013; Libert, 2015a). While the census literature frequently includes sub-analysis of sector-specific populations of web pages, there are relatively few studies which place significant emphasis on any particular socially-defined context.

While both approaches detailed above have produced a wealth of valuable findings, there remains scant accounting of the actual reach, impact, and policies of the actors collecting the bulk of user data in the OBA space. The companies which collect the vast volumes of data used for OBA know the deepest secrets of billions of users, but extant research approaches have insufficiently linked the normative dimensions of privacy values to the observed practices of specific companies. This situation has generated an information asymmetry by which companies know vastly more about the behavior of individuals than these individuals knows about the companies' behavior. Likewise, regulators and advocates acting on behalf of the public lack the data and computational resources enjoyed by companies in the OBA sector.

The only means to bridge the gap in literature and resolve the information

asymmetry between the public and the OBA industry is to employ methods which allow for grounding technical discovery within well-defined social-normative contexts. Such an approach facilitates linking the observed activities of companies in the OBA sector with the underlying social values which they potentially violate. Given the truly staggering reach of OBA, successful methods must also operate at a scale which is capable of fully documenting the behavior of billion-dollar data harvesting operations. As noted above, one of the primary contributions of this dissertation is the development of new tools which equalize the information differential between the public and the OBA industry, allowing new questions to be asked.

Methods: `webXray` and `policyXray`

Beyond the material contained in this text, an underlying contribution of this dissertation is the development of two new software tools. The first, `webXray`, facilitates massive-scale monitoring of data flows on the web and linking such flows to the companies which harvest user data. The second tool, `policyXray`, builds on `webXray` to provide automated auditing of privacy policies to determine if discovered data flows are appropriately disclosed to users. Prior to detailing the approaches used by these tools, it is necessary to briefly detail the technical mechanisms which power the online behavioral advertising (OBA) process.

A real-world example is the best way to understand how user information is leaked to third-parties on a typical web page. When a user searches online for "HIV"

one of the top results is for the U.S. Centers for Disease Control and Prevention (CDC) page with the address "http://www.cdc.gov/hiv/"[2]. Clicking on this result initiates what is known as a "first-party" Hypertext Transfer Protocol (HTTP) request to the CDC webserver (Figure 1.1). A portion of such a request is as follows:

```
GET /hiv/
Host: www.cdc.gov
User-Agent: Mozilla/5.0 (Macintosh...
```

This request is sent to the CDC webserver ("Host: www.cdc.gov") and is an instruction to return ("GET") the page with the address "/hiv/". This request also includes "User-Agent" information which tells the server what kind of browser and computer the user has. In this case, the user employs the Mozilla Firefox browser on a Macintosh computer. Such information is helpful when loading specially optimized pages for smartphones among other tasks.

Once this request has been made, the CDC webserver sends the user an HTML file. This file contains the text of the page as well as a set of instructions that tells the web browser how to download and style additional elements such as images. In order to get the CDC logo, the following HTTP request is made:

```
GET /TemplatePackage/images/cdcHeaderLogo.gif
```

[2]Note that findings in this section are current as of April, 2014 and the page has subsequently changed.

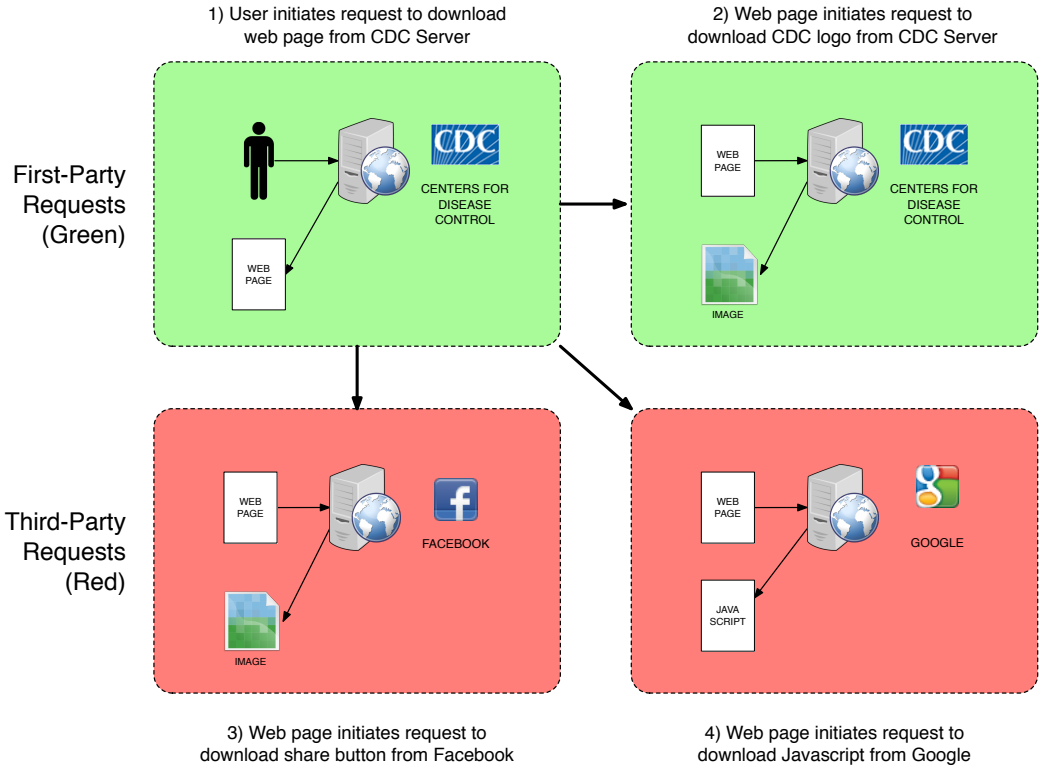

Figure 1.1: First- and Third-Party Requests on the CDC Web Page for HIV/AIDS

```
Host: www.cdc.gov

User-Agent: Mozilla/5.0 (Macintosh...

Referer: http://www.cdc.gov/hiv/
```

This request introduces a new piece of information called the "Referer" (a word which is misspelled in the original technical document and is in continuing use). The "Referer" contains the address of the page the user is currently viewing. The CDC web server may keep records of all HTTP requests in order to determine what pages and content are being requested most often.

Because the "Host" for both requests is identical (www.cdc.gov), the user is only interacting with a single party and such requests are called "first-party requests". The only two parties who know that the user is looking up information about HIV are the user and the CDC. However, the HTML file also contains code which makes requests to outside parties. These types of "third-party requests" typically download "third-party elements" such as images and Javascript.[3] This process happens in the background without user input, and the term "invisible web" is sometimes used to describe it.

On the CDC's HIV page, third-party requests are made to the servers of Facebook, Pinterest, Twitter, and Google. In the case of the first three companies, the requested elements are all social media buttons which allow for the sharing of

[3]Javascript is a programing language which, among other things, may be used to write programs which run inside a user's web browser.

content via the "Recommend", "Tweet", or "Pin It" icons. It is unlikely that many users would understand that the presence of these buttons indicates that their data is sent to these companies. In contrast, the Google elements on the page are entirely invisible and there is no Google logo present. One of these requests is sent to Google's Analytics service to download a file containing Javascript code:

```
GET /ga.js
Host: www.google-analytics.com
User-Agent: Mozilla/5.0 (Macintosh...
Referer: http://www.cdc.gov/hiv/
```

Again, the "Referer" field reveals that the user is visiting a page about HIV. By pairing information about the User-Agent, Referer, and user's IP address, it is possible for companies like Google and Facebook to identify people who are concerned with HIV (Yen et al., 2012). In all likelihood those visiting this page are unaware of this fact, and would not be happy to find out their personal data had been transferred without their consent.

Companies which are able to insert their elements into a given page may build behavioral profiles of users, thus a main goal of online advertisers is to get their elements on as many pages as possible. In many cases, advertisers pay to have their elements put on the page via revenue sharing arrangements, and this the dominant model in the publishing industry (Turow, 2012). Another way to get elements onto a page is to provide "free" services to web masters such as traffic analytics and social

media sharing buttons, as in the case in the CDC example. This is how Google, Facebook, and Twitter are able to amass huge volumes of information about users' habits once they leave their respective platforms and venture out onto the larger web - even on sites which do not rely on advertising revenue.

While many of the major OBA companies provide public accounting of their profits, and in some cases government requests for data, none of them provide any means to get a list of all the web pages from which they collect users' behavioral data. They provide virtually no public disclosure of the degree to which they collect data in sensitive contexts, and if they do, how it is treated. They provide limited means, at best, for users to discover tracking is taking place on the pages they are *currently viewing*. The structural disadvantages which allow companies to construct behavioral profiles of users while masking their own activities are rooted in the technical processes described above. Revealing these processes requires independent auditing of data flows.

`webXray` is a software platform developed to audit flows of user data on the web and to attribute such flows to the companies which collect user data. `webXray` is the most powerful tool of its kind and is capable of collecting data on a mass scale. Deployment of `webXray` has two primary components: choosing a population of web pages to analyze, and capturing data flows on such web pages in order to attribute web tracking behavior to the companies which practice it.

The studies contained herein use several methods for establishing relevant pop-

ulations of pages. In some cases, it is possible to model the pages a user would visit by utilizing search engines to produce pages related to specific terms, such as the name of a given disease. In other cases, lists of popular pages in a given category, such as newspapers, may be harvested. Finally, in cases where broad trends are of interest, popular site lists may be used to establish base-line populations.

Once a page population has been defined, the list of pages may be processed by `webXray`. The program uses a "headless" web browser (meaning it has no graphical user interface) to load pages, execute Javascript, set cookies, and record all network requests made for both first- and third-party elements. Once such data is gathered for each page, a database is constructed which contains all of the domains to which a user visiting the pages in the population would transmit data. An examination of these domains allows all requested elements to be classified as either first- or third-party, and reveals the underlying technical contours of data flows.

The collection of raw traffic data alone does not reveal the companies which receive user data, and by extension, the larger social forces at play. While it is possible to programmatically detect requests to third-party domains, it is not always clear who the requested domains belong to. However, by examining domain registration records, it is possible to pair seemingly-obscure domain names (e.g. "2mdn.net", "fbcdn.net") with their corporate owners (e.g. Google, Facebook). Over several years, `webXray` has incorporated an extensive library which links hundreds of domains to over 180 companies. By linking the domains which receive user data to

their corporate parents, it is possible to use network traffic as a means to reveal and interrogate the *social relationships* between users and companies. Quite often users are unaware of these relationships.

As noted above, the technical features of websites make it very difficult for users to know if, and when, their data is being transferred to third parties. `webXray` can reveal such connections, providing the basis for further inquiry. One type of such inquiry is determining if the privacy policy of a given page fully discloses the data flows to which users are subjected. This type of analysis has not been done before and `policyXray` is designed to accomplish this task.

When `webXray` loads a page it attempts to find a link to the site's privacy policy. Once `webXray` analysis is completed, `policyXray` may follow the link in order to extract the policy text. Once the policy is extracted, the names of the companies found in the `webXray` analysis are searched for in the policy text in order to determine if they are disclosed. In this way, it may be empirically determined at significant scale if users have reasonable means of being informed of their loss of their privacy, thus providing a comprehensive view of social, technical, and policy implications of data flows which drive the focus of this dissertation.

Structure and Chapters

This dissertation contains a review of the existing literature in privacy and online behavioral advertising, an extensive methods section, and three case studies. Each

of the case studies in grounded in specific social-normative contexts which provide a basis from which to both frame and explore technical findings.

The literature review chapter discusses larger theoretical approaches within the privacy studies field and then focuses on several broad areas of research: surveys which document public opinion towards privacy issues, qualitative work which sheds light on industry motivations, technical research which documents the methods and extent of online tracking mechanisms, and policy-oriented research. These branches of research all shed light on various aspects of the online behavioral advertising (OBA) phenomena from a range of methodological and theoretical viewpoints. The underlying contribution of this dissertation is highlighted by examining how extant social-normative, technical, and policy research issues have been addressed in a fractured, rather than holistic, means.

The methods chapter goes into great detail as to the novel and innovative design considerations taken with building socially-grounded populations of webpages, `webXray`, and `policyXray`. In regards to the page populations, several techniques are discussed which move the studies in this volume beyond the extant approach of looking at topic-related hubs to generating rich and expansive populations of web pages which reflect the complexity and nuance of social life online. The underlying architecture of `webXray` is discussed in-depth, along with coverage of how `webXray`'s approach to focusing on attribution of data flows to specific corporations represents a significant advancement over existing methodology. Furthermore, the flexibility

of `webXray`'s design is used to highlight how additional data sources related to malware analyses may be used. Finally, the methods used by `policyXray` to audit the disclosure of data flows, heretofore unattempted in the literature, are detailed.

The first case study, *On the Impossibility of Accepting the Unknown: A Web-Scale Analysis of the Failure of Notice and Choice*, traces the history of the Fair Information Practice Principles (FIPPs) to present day and provides the largest analysis to date of privacy policies online. The FIPPs are a set of normative principles originally conceived in 1973 which set out a basis from which to apply a rights-based approach to the governance of automated systems which process data on individuals. These principles expanded over time and have had profound impact on regulation around the globe. However, the online behavioral advertising industry has co-opted and weakened the FIPPs in order to construct a self-regulatory regime which fails to meet the original normative goals of the principles. The nature of the self-regulatory approach is a subject of significant controversy in academic literature, regulatory practice, and the public dialogue.

With this background and history in mind, the main methodological and substantive contribution of the study is an analysis of over 90 million data flows captured on one million websites, the privacy policies of over 200,000 sites, and the extended policies of the top 25 of 183 identified companies which receive user data. It is determined that not only are the self-regulatory principles promoted by industry a perversion of social norms, they are not followed in any meaningful way.

It is found that users are tracked by a huge array of companies, the majority of which do not provide consumer-oriented services, and are thus likely unknown by the vast majority of users. Likewise, an automated examination of the policies for over 200,000 sites demonstrates that only 10.7% of data flows are disclosed in site policies. The policies of web sites and the major collectors of user data are nearly impossible to understand based on the application of widely used readability metrics. Furthermore, industry policies ignore or reject user choices in data collection, and fail to provide minimum baseline security protections for data transfers. Overall, it is shown that the normative values embedded in Fair Information Practice Principles are being violated on a mass scale and the OBA industry's claims of self-regulation are proven to be a fiction.

The second case study, *Privacy Implications of Health Information Seeking on the Web*, grounds the failure of effective regulation in the OBA space within the universally recognized norms governing medical privacy. As with the prior case study, this one draws on social norms which predate OBA to highlight the problems associated with its use. Despite claims that privacy is a "modern" invention, the Hippocratic Oath, dating from the 5th Century B.C., has an explicit privacy clause and requires physicians to swear that *Whatever I see or hear in the lives of my patients...I will keep secret, as considering all such things to be private* (National Institutes of Health, History of Medicine Division, 2002). Today, regulations such as HIPAA uphold similar protections. The reasons for both the ancient and modern

commitments to medical privacy is that disclosure of patient information may lead to public embarrassment and discrimination. The impact of these harms is profound: patients who are scared of public disclosure of their medical conditions are less likely to seek care. Patients who do not seek care may spread disease and even die. While the impacts of losing medical privacy are profound, online advertisers take no oaths to protect patient privacy and collect vast volumes of medical-related data online.

One of the primary methodological contributions of the case study on medical privacy online is that it grounds the larger social context of medical privacy within a narrowly defined population of websites. While prior studies have analyzed a handful on topic-focused "hubs" of medical information, the Pew research center has found that the majority of Americans not only go online to seek health information but do so by using search engines instead of hubs (Fox and Duggan, 2013). The population of pages in this chapter is created by assembling a list of 2,000 common diseases, and using them as search terms in order to create a full population of over 80,000 pages related to health concerns. This population reflects the full variety of sources of online medical information ranging from newspaper articles, online support forums, hospitals, to non-profit and government entities.

An analysis of these pages reveals that users visiting them have their personal browsing habits subjected to third-party interception in over 90% of pages. Furthermore, in 70% of these cases, the precise name of a condition, symptom, or treatment

is revealed. An investigation of industry self-regulatory guidelines demonstrates that patient privacy is not protected by industry. More troubling, an evaluation of extant federal regulations and policies reveals that regulators are poorly equipped to defend users. Once again, it is shown that social norms are undermined by OBA and regulations are failing to rein in undesirable practices. OBA also threatens the health of democracies.

The third and final case study, *Watch Dog or Watch Tower? Privacy, Trust, and Security Implications of Behavioral Advertising on News Websites*, investigates the epicenter of the OBA phenomena: the online news sector. One of the primary sources of revenue for U.S. news media has long been advertising, and as news media have moved online, they have embraced OBA. However, in the transition from anonymous print advertising to the personalized targeting and surveillance of the OBA paradigm, the democratic function of the press has been deeply undermined.

The normative foundation of the American system of governance rests on a separation of powers between the executive, legislative, and judiciary branches. The reason for this division is that each branch acts as a check on the others, preventing the emergence of tyranny. In addition to these three formal branches of government, the U.S. Constitution provides broad press freedoms in order that news media may act as an *additional* check on power; the press is often referred to as "the fourth estate" for this reason.

While reality often differs from the ideal, the press is viewed as able to act as a

check on power for several reasons. First, the press is able to function due to the trust they have earned from citizens in being transparent and fair in their reporting. Second, the press operates from a position of independence and is ultimately beholden to the needs of citizens. Third, and most important, the press represents the locus of information and public deliberation about matters of governance; the ultimate check on power is an informed citizenry.

Online behavioral advertising compromises the above functions in several ways. First, the press is unable to function without the trust of citizens. However, the essential nature of OBA is premised on extracting user data in covert ways which survey data has demonstrated are rejected by the public. Furthermore, poor oversight of OBA facilitates the spread of criminal malware. Both covert data collection and the presence of malware undermine any basis of trust. Second, while press outlets require independence to operate without influence, OBA fosters a centralization of both revenue and content-delivery infrastructure, which gives a handful of OBA firms massive unseen leverage over the press. Finally, the free and open airing of facts and ideas which the press fosters is threatened by OBA as OBA provides a means for the government to surveil individuals, which may result in a chilling effect.

In order to explore the extent of OBA in the U.S. news media, a population of 250,000 pages drawn from over 5,000 news outlets has been assembled and analyzed with `webXray` and `policyXray`. The practices of OBA are widespread: nearly 98%

of news pages transfer user data to an average of 41.43 external domains, rates which far exceed what is normal for non-news websites. This transfer of user data is carried out using methods specifically designed to be hard to detect. An analysis of privacy policies shows that they are written in a manner which is unclear and only disclose 3.58% of data transfers taking place. Furthermore, news websites are not transparent in regard to the risks of OBA and 80.82% of pages expose users to domains which have delivered criminal malware. In both cases, there is little basis for trust in news websites.

In terms of press independence, a small group of companies control the underlying infrastructure which drives OBA on news websites, putting revenue in the hands of powerful forces outside the control of publishers. This centralization of revenue means that the interests of the OBA industry may come before allegiance to the public. Furthermore, an analysis of underlying data flows reveals that publishers increasingly lack the ability to host their own websites without the help of OBA firms; this dynamic reduces independent publishers to mere content providers.

Finally, not only has the spread of OBA eliminated any semblance of privacy, an analysis of leaked N.S.A. documents reveals that the presence of OBA on news websites may allow the government to monitor citizens as they read the news. This raises the potential that individuals may be fearful of expressing controversial ideas or seeking out information. Such an impact could result in a chilling effect of the democratic participation. The intrusion of government surveillance inverts the fun-

damental role of the press as a check on, rather than facilitator of, power. Overall, the impact of OBA is to weaken the ability of the press to fulfill its democratic function.

The final chapter ties together the findings of the three case studies in order to advance a larger argument of how OBA represents a fundamentally socially corrosive practice which undermines norms, violates personal dignity, and poses an existential threat to the future of free societies. It further addresses questions of how the situation may be fixed.

The issue of self-regulation is explored in order to create a portrait of what exactly an effective program of "notice and choice" would look like. Based on the huge volumes of data analyzed in the three case studies, it is found that following a true notice and choice system would result in such a deluge of notifications that it would render the web unusable. Furthermore, an investigation of the normative characteristics of both medical privacy and the value of the press illustrate that in many cases a "choice" in favor of OBA is in fact a choice to sacrifice vital social norms which benefit all of society in favor of a revenue model which favors a handful of technology executives.

The evidence presented herein provides proof that self-regulation is failing, real social harms are occurring, and it is time to rapidly adopt a regulatory stance which is rooted in social norms. For regulatory approaches to be successful, they must adopt the type of advanced technological methods used in this study. While the

overall findings of this dissertation raise cause for concern, by pairing normative concerns and large-scale data analysis it is possible to contain and regulate the OBA sector. Both the normative framings and technological tools needed for this task are used in the following chapters.

Chapter 2

Literature Review

Given the increasingly outsized role the web, mobile applications, and other Internet services have in the essential organization of social life, there is widespread concern over how information about users of these services is collected, stored, and processed. Scholars from a number of academic disciplines have addressed the issue of online data transfer and shed light on a variety of factors impacting user privacy. Given that this dissertation focuses on the topic of privacy on websites, the relevant literature is firmly tied to this specific aspect of online data transfer.

Despite the fact that the topic area itself is fairly well defined, the literature remains diffuse and poorly organized. Existing disciplinary boundaries have created several intellectual ghettos between computer science, communications research, law, and public opinion research. Although there are many examples of impressive cross-disciplinary projects, such research tends to be the exception rather than the

rule. A major purpose of this dissertation is therefore to weave together these many strands of scholarship in order to promote a holistic understanding of online privacy. This is done by developing new computational methods specifically designed to address social, rather than purely technical, questions.

This chapter details some of the most notable literature in the areas of theory, public opinion, qualitative industry research, and computer science. Additional chapters contain topic-specific reviews related to privacy regulation, medical privacy, and normative concerns related to the press.

Theoretical Foundations: Moving Beyond Dichotomy

Privacy is often treated as a zero-sum game between two opposing poles: the "private" and the "public". In this conception, information often originates in a form only known to a single or closed group of persons, at which point it is "private". Then, through some compromise or negotiation, the information becomes known to outsiders and is thus rendered "public". The interests of individuals and society are thus situated between two opposing poles.

This polar approach has been used with success to frame many issues. For example, Alan Westin has demonstrated that privacy may be understood as a means to explicate differences between democratic and non-democratic societies (Westin, 1967). However, as with any dichotomous system, nuance can be lost. Helen Nissenbuam's theory of "Contextual Integrity" demonstrates how using theories

which move beyond dichotomy allows for deeper interrogation of the subject matter (Nissenbaum, 2004, 2010).

If, as it seemed, the new technology was on a collision course with the values of personal privacy and human dignity, could the collision be averted? Could a system be devised to identify and permit beneficial uses of the new technology and yet, at the same time, preclude those uses that most men would deem intolerable?

- Alan Westin, 1967

Written in the early days of large-scale computerized databases, Alan Westin's 1967 book, *Privacy and Freedom*, provides a prescient view on the relationship between privacy and democratic norms and personal freedom. At the most basic level, Westin defines privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" (Westin, 1967). Within the broader social context, "privacy is the voluntary and temporary withdrawal of a person from the general society through physical or psychological means" (Westin, 1967). Thus, in Westin's formulation, privacy is positioned as a barrier between the individual and the group.

Altering the privacy barrier has drastic effects upon systems of governance. On one extreme, the denial of privacy to individuals is a mark of non-democratic, authoritarian regimes. According to Westin, "The modern totalitarian state relies on secrecy for the regime, but high surveillance and disclosure for all other groups" (Westin, 1967). In this view, the government which most strongly opposes privacy

protections for citizens does so "to foster a sense of loneliness and isolation", which results in "identification with the state" (Westin, 1967). Westin makes a strong argument that a repressive government may gain strength by limiting privacy.

On the other extreme is the egalitarian democratic society, in which the protection of privacy, rather than the prohibition on it, provides institutional strength. Westin asserts that "The democratic society relies on publicity as a control over government, and on privacy as a shield for group and individual life" (Westin, 1967). One aspect of this is to defend the right of individuals to have group associations as groups "provide opportunities for sociability, expression of independent ideas, resolution of community conflicts, criticism of government, and formation of a consensus on public policy" (Westin, 1967). Likewise, in the realm of personal life, the ultimate expression of democratic participation is the vote, which as Westin notes, is performed by secret ballot. In this way, privacy of the group and the citizen strengthen both the maintenance and expression of democracies.

While Westin's approach has great value, advances in consumer computing technologies have vastly complicated the fundamental framing of his theory as privacy as a barrier between personal and political life. This can be seen most clearly in the ongoing debate over government surveillance. Whereas Westin's theory is based on a dichotomous separation of citizen and government, intelligence agencies now rely on private corporations to function as an intermediary layer in a "surveillant assemblage"(Haggerty and Ericson, 2000). This shift illustrates the limitations of

dichotomous approaches.

To overcome the limitations of the public/private dichotomy, Helen Nissenbuam has proposed a theoretical framework she calls "Contextual Integrity". This framework treats privacy not as a static concept related to the individual, but as a dynamic concept in which structured social relationships among various actors determine how information should be transferred. The value of contextual integrity is that it "ties adequate protection for privacy to norms of specific contexts, demanding that information gathering and dissemination be appropriate to that context and obey the governing norms of distribution within it" (Nissenbaum, 2004).

Prior to delving into the specifics of Nissenbaum's theory, it is instructive to consider three traditional privacy framings which she has determined to be inadequate. The first is "Protecting Privacy of Individuals Against Intrusive Government Agents" (Nissenbaum, 2004). This framing draws from Westin's conception in which privacy serves as a means to achieve parity between the individual and the state. The second privacy principle is "Restricting Access to Intimate, Sensitive, or Confidential Information", which is a means to protect classes of information which are posited to have intrinsically privacy-sensitive content; Nissenbuam provides examples of regulations specific to health, education, and finance (Nissenbaum, 2004). The third and final principle, "Curtailing Intrusions into Spaces or Spheres Deemed Private or Personal", is a reformulation of the classic "castle doctrine" in which a person's property and home are considered the barrier between the private and the

public spheres (Nissenbaum, 2004).

While the preceding three principles cover much ground in the privacy debates, they all promote a dichotomous zero-sum conception. One area where this is most limiting is to be found in what Nissenbaum calls "public surveillance". In this case small pieces of otherwise public information, such as a person's location as observed by strangers on the street, may be aggregated to reveal "private" issues, such as a visit to a doctor's office, place or worship, or engagement in activities which may be frowned upon by prevailing social norms. [1]

Nissenbaum points out that public surveillance is not sufficiently captured by any of the three traditional privacy principles in that it "does not involve government agents...collection or disclosure of sensitive, confidential, or personal information; or intrusion into spaces or spheres normally judged to be private or personal" (Nissenbaum, 2004). Likewise, when records of visits to public websites are aggregated, many pieces of seemingly anonymous information may be combined into a sensitive whole.

Nissenbaum's solution to the problem of dichotomy is to focus on contexts, which she defines as "structured social settings characterized by canonical activities, roles, relationships, power structures, norms (or rules), and internal values (goals, ends, purposes)" (Nissenbaum, 2010). The transfer of information within and between contexts is termed a "flow" and Nissenbuam asserts that "there are no arenas of

[1] Supreme Court Justice Sonya Sotomayer's concurring opinion provides an excellent discussion of issues related to location data in *US v. Jones* (Supreme Court, 2011)

life *not* governed by *norms of information flow*, no information or spheres of life for which "anything goes" (Nissenbaum, 2004).[2] Combining these two concepts, it may be seen that the core focus of the framework is to develop and interrogate "context-relative informational norms" (Nissenbaum, 2010).

Informational norms are constituted by two factors, termed "appropriateness" and "flow". Nissenbaum suggests that violations of contextual integrity are found when either of these norms are contravened (Nissenbaum, 2004). According to the framework, "norms of appropriateness dictate what information about persons is appropriate, or fitting, to reveal in a particular context" (Nissenbaum, 2004). An example of "appropriateness" is found in how detailed information about one's health is appropriate to discuss at a doctor's office, but it may be illegal for a potential employer to inquire on such matters.

Norms of "flow" may generally be understood as the process by which information from one context enters another via a means of information distribution conducted by actors other than the subject. An example of an unacceptable information flow would be if a doctor authored a Facebook post revealing the condition of a patient to numerous outside actors.

Context-relative information norms emerge from four dimensions which Nissenbuam calls "contexts, actors, attributes, and transmission principles" (Nissenbaum, 2010). *Contexts* are socially defined; *Actors* are defined as "senders of information,

[2]Emphasis in original.

recipients of information, and information subjects" (Nissenbaum, 2010); *Data Attributes* are the type of information which is being exchanged; the *Transmission Principle* dimension is another term for "policy" and is defined as "a constraint on the flow (distribution, dissemination, transmission) of information from party to party in a context" (Nissenbaum, 2010).

Survey and Qualitative Literature

In the context of web browsing there are generally two broad sets of actors whose lives are impacted by information flows. The first is comprised of the general public who use the Internet. There is a broad and deep public opinion literature detailing public concerns with online information flows. The second group of actors are those receiving the data: online advertising companies. Qualitative and interview research gives the best insights into the motivations of the advertising industry.

Within a democratic context, the opinions and attitudes of informed citizens on topics of governance should provide guidance to policy makers; likewise, the feelings and values of those subjected to data collection should inform transmission principles. Given that in the United States, privacy regulation is constituted as a mix of government and industry efforts[3], democratic-normative approaches towards privacy research seek to quantify public opinion as it relates to both generalized privacy values as well as specific data collection practices. There is a large volume

[3] "Effort" on industry side often being an effort *against* effective regulation.

of high-quality public opinion research on the topic stretching back to the earliest days of the web.

Concerns about privacy on the web have been documented for nearly two decades. A 1999 survey determined that "respondents registered a high level of concern about privacy in general and on the Internet" (Ackerman et al., 1999). Likewise, a 2003 study found that "a clear majority of Americans express worry about their personal information on the web" (Turow, 2003). Giving credence to the theory that public opinion reflects democratic-normative goals, a 2006 study found that privacy policies were "considered a highly important issue in the public debate by our respondents" (Acquisti and Gross, 2006).

A 2009 survey found that "69% of American adults fe[lt] there should be a law that gives people the right to know everything that a website knows about them" (Turow et al., 2009). Similar work in 2012 showed that 60% of respondents would like the proposed "Do Not Track" (DNT) standard to prevent websites from collecting personal information (Hoofnagle et al., 2012). In a 2014 study, the Pew Research Center discovered that 70% of respondents felt records of the websites they had visited constituted "very" or "somewhat" sensitive information (Madden, 2014). The general trend established by these surveys is not just that most users value privacy, but that the current state of privacy online represents an area of significant anxiety.

The studies mentioned above are all high-quality empirical academic works; how-

ever, it is important to note that within the literature of privacy and public opinion there is a large volume of deceptive studies. Such studies are usually produced by industry groups who misuse survey methodology in order to falsely generate "evidence" in support of industry positions. To see how this works, first consider two different survey questions: "If companies give me a discount, it is a fair exchange for them to collect information about me without my knowing it.", and "Which of the following would you prefer: an Internet where there are no ads, but you would pay for most content like blogs, entertainment sites, video contact and social media, or today's Internet model in which there are ads, but most content is free?". Each of these questions was administered in survey form.

Ostensibly both questions are framing core aspects of contemporary online advertising practices (whereby "free" content is given as a result of hidden data transfers), yet each question produces very different conclusions. The first question, administered by academic researchers, yielded the conclusion that "a large pool of Americans feel resigned to the inevitability of surveillance and the power of marketers to harvest their data" (Turow et al., 2015), whereas the second survey, administered by industry group Digital Advertising Alliance, found that "Americans place great value on the availability of free Internet content, and appreciate Internet advertising that is tailored to their specific interests" (Digital Advertising Alliance, 2013). Not coincidentally, the conclusion of the second survey supports an industry position. Sadly, this is not an isolated example, but part of a pattern which the

industry has used to misrepresent public opinion in their favor. This practice has been well documented by Gandy who states that "In the past 25 years, references to public opinion have been used to frame the public as...willing to negotiate their privacy demands" (Gandy, 2003).

In sum, there is a rich literature of valuable public opinion research conducted in an academically rigorous and empirically valid manner. This literature has consistently supported the conclusion that privacy is highly valued in the abstract, and specific data collection practices are widely reviled. However, when utilizing public opinion research it is important to maintain awareness of literature which is of poor quality and performed for duplicitous ends.

While survey work is indispensable for public opinion research, survey instruments are an awkward fit for understanding the motivations and values of the second set of "actors" present in web privacy: the online advertising industry. An approach to gain insights into this world is to employ observational methods, taking cues from the fields of anthropology and sociology. Such observation may be carried out both by close reading of primary source documents such as trade publications, as well as through in-person interviews. Such an approach allows for scholars to develope nuanced portraits of those on the receiving end of online data flows.

In *Niche Envy* and *The Daily You*, Joseph Turow investigates the data broker and online advertising industries and seeks to address the question "What is the social significance of executives' newfound insistence that consumers are morally

obligated to pay attention to advertisements in return for the free or discounted media material they receive?" (Turow, 2008). Left out of this question, but a major focus of both books, is the question of how the data broker industry justifies practices which survey research demonstrates consumers find objectionable. To answer this question, Turow relies on a close reading of the trade press and interviews with industry insiders.[4]

According to Turow, trade press "literature has helped to shape [his] understanding of the basic contours of the media-and-marketing system" (Turow, 2008). The pages of the trade press give a window into decades of anxiety over topics such as online comparison shopping and the role of advertising and the family. Likewise, Turow has attended industry events to see debates from the trade press taking place in person. Demonstrating that listening to the language used by industry gives deep insights into the social values of companies, he has discovered that the industry uses callous terms such as "target" and "waste" to describe the difference between desirable and undesirable consumers (Turow, 2012).

Overall, qualitative methods allow researchers to tell the story of a particular set of actors in such a way that their value system and motivations may be understood by outsiders. The next layer of literature moves from the motivations of these actors to documenting their actions.

[4]In this context, "trade press" is primarily periodicals such as *Advertising Age* which are written for a narrow audience of practicing marketing professionals.

Technical Literature

As noted in the previous chapter, OBA has fostered a massively profitable industry which invests significant resources into collecting as much data as possible about user behavior using constantly evolving techniques. For this reason, detecting the technical mechanisms powering OBA is of significant and sustained interest to to a wide array of computer security researchers. These researchers have produced a rich strand of literature devoted to understanding the technical underpinnings and privacy threats posed by OBA.

The technical literature regarding OBA and data attributes is largely broken up along two major themes: analyzing the functioning of specific types of computer code delivered to users' computers (called the "client-side") and censuses of the reach of online advertising networks.

Much of the activities powering OBA, such as creating mathematical models which predict a user's propensity to click a given advertisement, take place on the servers of advertisers, are protected as intellectual property, and are exceedingly difficult to study.[5] However, in many cases, online advertisers deliver code to users' computers which manipulates the computer to deliver extra information about users' behavior, allowing users to be more easily "tracked". Thus, researchers may both investigate the characteristics of clients which allow them to be exploited

[5]Despite these difficulties, there are some early promising efforts (Sweeney, 2013; Lécuyer et al., 2014)

as well as parties which are leveraging such exploits.

Research into novel client-side tracking techniques is an ongoing project, with one early paper, "Timing Attacks on Web Privacy" by Felten and Schneider, appearing in 2000. In this paper, the authors describe a means of exploiting the caching properties of web browsers in order to determine if a user has previously visited a given website. With foresight, they further observe that such an attack could be enacted if "A Web advertising agency could add measurement code to the banner ads it distributes" (Felten and Schneider, 2000).[6] This research was followed up by Jackson et al. in 2006, with the finding that "a variety of means, including...browser cache methods and inspecting the color of a visited hyper- links...can be exploited to track users against their wishes" (Jackson et al., 2006).

One of the most innovative and well-known studies, "How Unique Is Your Web Browser?" by Peter Eckersley, begins with the observation that "The most common way to track web browsers...is via HTTP cookies, often set by...3rd party analytics and advertising domains" (Eckersley, 2010). However, the paper identifies a more invasive, and durable form of tracking: browsing fingerprinting. As with the physical variety, browser fingerprints are generated from combinations of otherwise innocuous factors which when combined create a unique identifier. While a detective may use the collection of loops and whorls on a finger, a browser fingerprinting

[6]As a testament to the age of Felten and Scheider's study, they performed their study on a "Netscape Navigator 4.5 browser on a Windows NT 4.0 (Service Pack 4) PC with a 350 MHz Pentium II processor and 256 Megabytes of RAM".

algorithm may use the combination of User-Agent, Screen resolution, timezone, browser plugins, and system fonts to uniquely identify a given browser (Eckersley, 2010). Indeed, Acar et al. created "a framework for the detection and analysis of web-based fingerprinters" called "FPDetective" (Acar et al., 2013), Nikiforakis et al. discovered evidence that companies have employed "the circumvention of HTTP proxies to discover a user's real IP address" (Nikiforakis et al., 2013), and Jang et al. have investigated various uses of malicious Javascript in the wild (Jang et al., 2010).

As noted above, it is often difficult, if not impossible, to determine what happens to fingerprints gathered on the client-side when they are processed on the servers of advertising networks. However, a study by Yen et al., "Host Fingerprinting and Tracking on the Web: Privacy and Security Implications", had access to data from the Microsoft corporation which included information on "millions of hosts across the global IP address space" (Yen et al., 2012). In this rare glimpse behind the curtain of web tracking, the researchers revealed that solely relying on data found in basic HTTP requests (User-Agent and IP addresses) they were able to identify unique browsers with 80% accuracy, which is in par with cookies (Yen et al., 2012). This important study suggests that even without robust browsing fingerprinting mechanisms, simple HTTP requests are sufficient to track users with high accuracy.

The second general area of technical research focus has been in censuses. In this case, the interest of researchers is in selecting a population of websites and con-

ducting a count of various forms of tracking mechanisms, third-party requests, and data flows. Much of the pioneering work in this area has been conducted by Krishnamurthy and Wills. In 2006 they advanced the concept of the "privacy footprint" which they describe as a means to analyze the "diffusion of information about a user's actions by measuring the number of associations between visible nodes via one or more common hidden nodes" (Krishnamurthy and Wills, 2006). What Krishnamurthy and Wills describe as "hidden nodes", may be more widely included under the umbrella of third-party HTTP requests. This work was followed up several years later with additional longitudinal studies (Krishnamurthy and Wills, 2009; Krishnamurthy et al., 2011).

While Krishnamurthy and Wills have done an excellent job in the area of measurement, they are not the only researchers exploring the topic. Castelluccia and colleagues have recently analyzed "the flows of personal data at a global level" (Castelluccia et al., 2013), and found intriguing differences in the nature, scope, and ownership of tracking mechanisms around the world. Other researchers have developed desktop browser add-ons (Roesner et al., 2012; Mayer and Mitchell, 2012) which they have used to detect the presence of a variety of tracking mechanisms. Some of the most recent measurement literature focuses on fingerprinting specifically (Acar et al., 2013). Additional research has revealed extensive tracking on websites containing health information (Libert, 2015b) as well as large-scale analysis of the both tracking and policies on the top one million Alexa websites (Libert,

2015a; Englehardt and Narayanan, 2016).

A common theme among all measurement research is that the amount of tracking on the web is increasing, and shows no signs of abating. However, this literature often provides a weak basis for tying technical findings back to the social norms which are being contravened.

Conclusion

While there is a fair amount of cross-pollination between the bodies of literature cited above, underlying disciplinary and methodological considerations have resulted in lost opportunities to create a holistic understanding of privacy concerns on the web. Largely this is because quantitative technical research has been poorly connected to social-normative foundations and vice-versa.

Researchers grounded in social norms often lose opportunities to strengthen their arguments through the use of empirical data, primarily operating in the abstract. Conversely, technical researchers focusing on client-side techniques have established a trend by which they discover the new technique-du-jour of the OBA industry without shedding much light on the larger trajectory and social meaning of the practices observed. Lastly, and most important to this dissertation, sensitive contexts of online data transfer are poorly defined in the web privacy measurement literature, pushing much research back into the trap of the decontextualized public/private dichotomy.

The major contribution to the field of privacy research made by this dissertation is that the new tools developed for the studies contained herein draw from the strengths of computer science in order to provide a much deeper insights into the questions raised by normative theorists. This approach allows social-normative contexts to be defined in a way which is consonant with the theories of Nissenbaum and tied to specific web pages. It allows data flows to be examined using technical methods which are capable of operating at scale equivalent with the collectors of data. Finally, by providing clarity as to how observed practices violate social norms, it provides a basis for new approaches to regulation.

Chapter 3

Methodology

As noted in Chapter 1, the purpose of this dissertation is investigate how the use of online behavioral advertising (OBA) has produced negative outcomes in specific social contexts. No methods existed for conducting this type of study, thus new methods were developed for the task and are described in this chapter. These methods must meet several goals concurrently. First, the very presence of third-party data collection on the web must be empirically proven. This involves three elements: selecting populations of web pages to analyze, detecting flows of user information on these pages, and tying these information flows to the companies receiving data. Second, specific social norms, policies, and regulations must be identified in order to ground the empirical findings in social-normative contexts. Third, a convincing link must be drawn between data flows on the web, violations of established social norms, and the policy failures which allowed them to happen.

This chapter first provides background on novel techniques for building contextually focused populations of web pages. This is followed by a description of a software platform, `webXray`, which has been purpose-built to collect data on third-party data flows at the network level and provide analysis of the parties receiving data. A related software tool, `policyXray`, has been developed to perform the previously unattempted task of auditing website privacy policies at large scale and is also described. Additionally, methods for evaluating malware exposure and generating relative measures between page populations are described. The limitations of these tools are discussed as well.

Context on the Web: Building Page Populations

As noted above, the purpose of this dissertation is to ground OBA practices in established social norms. In the technological context of the web, the types of pages a user is visiting is directly reflective of, and therefore determines, social context and norms.

If one considers the act of browsing the web to be a de facto private context, all web pages in the world would present a universal relevance to the investigation of privacy. However, such an open-ended approach fails to direct attention to the most vital social issues. It is more helpful to consider that within the totality of web pages, some have more privacy salience than others, and this salience is determined by how the pages connect to social norms which are external to the act of web

browsing.

In order to map web pages to external social contexts, specific areas of interest and populations of relevant pages must be defined. Once a population of pages is built on a solid theoretical and normative base, data flows may be interrogated and understood. There are two main ways to build these populations: the first is modeling the types of pages a user would view if searching for information related to a given topic, and the second is using available classifications of pages as the basis for examination.

While topic-based "portals" and "hubs" were common ways to start a search for information in the early days of the web, search engines have taken primacy. For example, on the topic of health, the Pew Research Center has determined that among the 72% of adult Internet users who went online to learn about health conditions, 77% skipped portals and instead "began at a search engine such as Google, Bing, or Yahoo" (Fox and Duggan, 2013). In order to construct a population of pages which is tied to a given context, it is helpful to replicate user interaction with a search engine by selecting pages based on search results. This requires two steps: first, selecting terms related to a context, and second, harvesting search results.

When using a search-based approach to building a page population, the selection of search terms is of vital importance: they need to be broad enough to capture a range of web pages a user would visit, but also narrow enough as to exclude ir-

relevant pages. It is important to acknowledge that a real part of lived experience on the web is to visit some irrelevant pages in search for the information the user desires, thus the standard of relevance is not topic-relevance per se, but relevance to what pages a user may visit. Likewise, depending on topic, *pages* relevant to a specific topic may be found on *sites* which are not ostensibly topic-related. For example, on a discussion forum of a close-knit community of gardening enthusiasts, it is possible to find threads in a "general discussion" sub-forum dealing with unrelated issues, such as health conditions.

The chapter on health privacy in this dissertation illustrates a search-based approach. A list of 2,000 common health conditions was harvested from the Centers for Disease Control, the Mayo Clinic, and Wikipedia. If a patient is given a specific diagnosis from a doctor, it is highly likely she or he will search for specific health-related terms and visit the top results which are returned from a search engine, likely going several pages deep into the results. For the health chapter it was determined that the first 50 search results for each term (corresponding to five pages of results), would be appropriate.

Once search terms are specified, the remainder of the task is purely technical. Given that Google is the world's most popular search engine, it would appear that submitting search terms to Google would be most effective. Unfortunately, Google does not offer programmatic API access to their search results and prohibits scraping, making the task difficult.[1]

[1]It is also worth noting that because Google tailors search results based on inferred attributes

The second place search engine, Microsoft's Bing, is amenable to research tasks.[2] While Bing's share of web search is a fraction of Google's, due to corporate partnerships, Bing results are found in Yahoo! Search, DuckDuckGo, and Apple's Siri virtual assistant (Yahoo!, 2015; DuckDuckGo, 2016; Apple, 2013). Microsoft offers an Application Programming Interface (API) by which Bing data may be retrieved by computer programs.[3] Furthermore, the API returns results based on a specified language and region, ensuring that results are simultaneously more generalizable as well as more targeted.[4]

For the chapters in this dissertation using a search-based approach to constructing page populations, a simple Python program is used to send search terms to the Bing API and retrive results. This approach generates list of pages which may come from the same site; thus it is important to underscore that under this approach aggregate measures are on a per-page rather than per-site basis. For example, if 20% of pages in a search-based population feature a given third-party element this may be due to the fact that a single site uses that tracker on many pages, even if other sites do not use it.

While search-based populations are very good at reflecting the experiences of

or prior activities of users, it is impossible to say if results gained from scraping would be generalizable.

[2]http://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/

[3]https://datamarket.azure.com/dataset/5BA839F1-12CE-4CCE-BF57-A49D98D29A44

[4]For example, to get English results for the U.S. the API call may be appended with the string "Market=en-US")

web users as they search for information, they are not always an optimal way of establishing context. For example, if a relevant context is culturally or nationally defined, it is often helpful to study the most popular websites in that context - regardless of their relationship to specific search terms. Likewise, in the context of news seeking in the United States, the top 50 Bing results for the term "newspaper" would fail to generate a page population which reflected the specified contextual attributes. In these situations, indices of sites sorted by topic and ranked by popularity are most useful. The Amazon subsidiary Alexa provides such lists.

Alexa offers three main types of indices according to global rankings, country rankings, and category rankings. These rankings are based on "a sample of millions of Internet users using one of over 25,000 different browser extensions" and "direct sources in the form of sites that have chosen to install the Alexa script on their site and certify their metrics".[5] The list of global ranks (which contains one million websites) "is a measure of how a website is doing relative to all other sites on the web over the past 3 months"; the list of country-specific rankings is "is a measurement of how a website ranks in a particular country relative to other sites over the past month".[6]

In terms of scientific validity, Alexa's methodology leaves much to be desired as it may be plagued by sample bias. However, considering that traffic data is highly guarded by publishers, there are few alternatives and Alexa represents a well-

[5]http://www.alexa.com/about

[6]http://www.alexa.com/about

established source for academic research (Krishnamurthy and Wills, 2006; Krishnamurthy et al., 2011; Roesner et al., 2012; Libert, 2015a; Englehardt and Narayanan, 2016).

Alexa also provides site ranking data according to category. The top-level categories supplied by Alexa closely mirror those of the now-struggling DMOZ project which claims to be "the largest, most comprehensive human-edited directory of the Web...historically known as the Open Directory Project (ODP)"[7].[8] In cases where a specified privacy context lines up with Alexa's available categories, constructing a population is as simple as harvesting links on a per-category basis.

It is also possible to search for relevant pages within indexes of category-specific sites. For the chapter covering political information on U.S.-based newspapers, the term "politics" was searched for within the set of sites in the "News - Newspapers - Regional - United States" category in Alexa and DMOZ. This provides a population of pages which is both targeted and circumscribed by sector.

As noted at the start of this section, matching a population of pages to a given social context is the vital first step which grounds technical findings in a social-theoretical framework. However, in order to explore where violations of social norms are in play, it is necessary discover the invisible data flows which reveal user actions to external actors.

[7]http://www.dmoz.org/docs/en/about.html

[8]DMOZ in turn was based on USENET.

Third-Party HTTP Request Collection and Analysis with `webXray`

There are two main steps which must occur once a page population has been defined: detecting information flows at the network level, and analyzing these flows in order to situate them in a social-normative context.

A primary goal of this dissertation is to investigate data flows on the web which expose users to third-parties of whom they may not be aware. On the web such data flows may be measured by examining Hyper Text Transfer Protocol (HTTP) network requests. Simply put, when a page is loaded in a web browser, it contains instructions which tell the browser to make HTTP requests to download various types of content ranging from images and videos to fonts and Javascript code.

Some HTTP requests are made to the same address which the web page is hosted on, called the first-party. A user may often visit the privacy policy of the page she or he is visiting with relative ease. However, additional requests may be made to actors external to website being visited, these are known as "third-party HTTP requests". Due to the fact that users may often be unaware of third-parties, and are often not exposed to their privacy policies directly, such requests may negatively impact privacy.

In order to investigate data flows it is first necessary to capture all HTTP requests made when loading a web page and separate first- and third-party requests.

One way to solve this problem is to simply evaluate the source code of the page. However, there are two key problems with this approach. First, elements such as "iframe" may include additional source code which would generate more requests, and this may be missed. Second, it is quite common for element addresses to be intentionally obfuscated in Javascript source code, making it impossible to parse them with standard libraries[9].[10]

Considering that parsing HTML is not sufficient to solve the problem, the only viable route is to execute the code in a web browser. The major advantage of using a browser is that all elements may be fully loaded, embedded Javascript code may be run, and the network traffic will reflect real-world conditions. One common way to perform this type of analysis is to use a browser automation tool such as Selenium to load pages, and then use a network proxy or browser add-on to catalogue HTTP requests.[11] One such add-on is "FourthParty"; however, at time of writing it is not compatible with the current version of Firefox and has not been updated in nearly a year (FourthParty, 2013).

Despite the lack of updates for FourthParty, the OpenWPM project uses this ap-

[9]The most common technique is to manipulate the Javascript "document.write" method to insert more Javascript.

[10]While sometimes benign, obfuscation techniques may also be used to complicate detection. Arthur C. Clarke famously said "Any sufficiently advanced technology is indistinguishable from magic."; I would add to this, "Any sufficiently advanced form of online advertising is indistinguishable from malware."

[11]http://docs.seleniumhq.org

proach in order to drive a full version of Firefox with Selenium[12] A major drawback is that this approach requires running a graphical user interface (GUI), which has material cost in the form of higher computing power. Furthermore, neither Fourth-Party nor OpenWPM correctly attributes network traffic to the actors which receive user data, providing an incomplete and decontextualized view of the data.

Having established parsing HMTL is insufficient, and leveraging a full web browser is fraught with complication, a third approach to the problem has been found: headless automation. In this context, "headless" refers to a web browser which does not require a GUI, but instead runs in a text-based command line environment. One of the earliest, and most well-known, headless browsers is PhantomJS. According to the official website, "PhantomJS is a headless WebKit scriptable [browser] with a Javascript API" and is an "optimal solution" for network monitoring (PhantomJS, 2016).[13] PhantomJS uses the WebKit engine which currently powers Apple's Safari browser and is the progenitor of Google's Chrome browser. Thus, while PhantomJS is not identical to using a Selenium-driven desktop browser, the underlying rendering engine is an effective replacement for major browsers.

[12]https://github.com/citp/OpenWPM

[13]As a testament to the influence of PhantomJS, newer headless browsers such as SlimerJS and trifleJS implement the PhantomJS API to run on top of the Firefox and Internet Explorer browser engines. PhantomJS has also inspired a small cottage industry of open-source projects with supernaturally-themed names: CasperJS, SlimerJS, ZombieJS, Ghost Driver, Poltergeist, SpookyJS, and more.

The most important aspect of PhantomJS is that it is a "scriptable", meaning short computer programs may be written which control the browser via an Application Programming Interface (API). The PhantomJS API has numerous features which may be used for development tasks. Most useful for this project are the features which allow for monitoring network traffic, thereby revealing the actors users are exposed to when loading a given page.

Now that the means to construct context-relevant page populations and extract HTTP requests have been established, it is time to describe a system for automating the process of collecting and analyzing data flows in a specified population: this system is called `webXray`.

`webXray` is written primarily in Python and uses a modular object-oriented architecture which provides strong separation between the browser engine, collection mechanism, database, and analysis mechanism. While only a single browser engine (PhantomJS) and database engine (MySQL) are supported at present, due to the engine-specific code being segregated in appropriate classes (PhantomDriver and MySQLDriver), it is fairly easy to adopt new browser and database engines as desired.[14]

The centerpiece of the program is a command-line interactive program which manages the selection and verification of page addresses and generation of reports. Database functions and management are purposefully hidden from the user in order

[14]Exploratory tests with replacing the PhantomJS engine with the Firefox-based SlimerJS engine were successful but require greater testing.

to facilitate ease-of-use. The program has command-line flags and may be run in an interactive mode ("-i"), unattended modes for collecting ("-c") and analyzing ("-a") large amounts of data, as well as a "single" option ("-s") which facilitates a quick analysis of a single page.

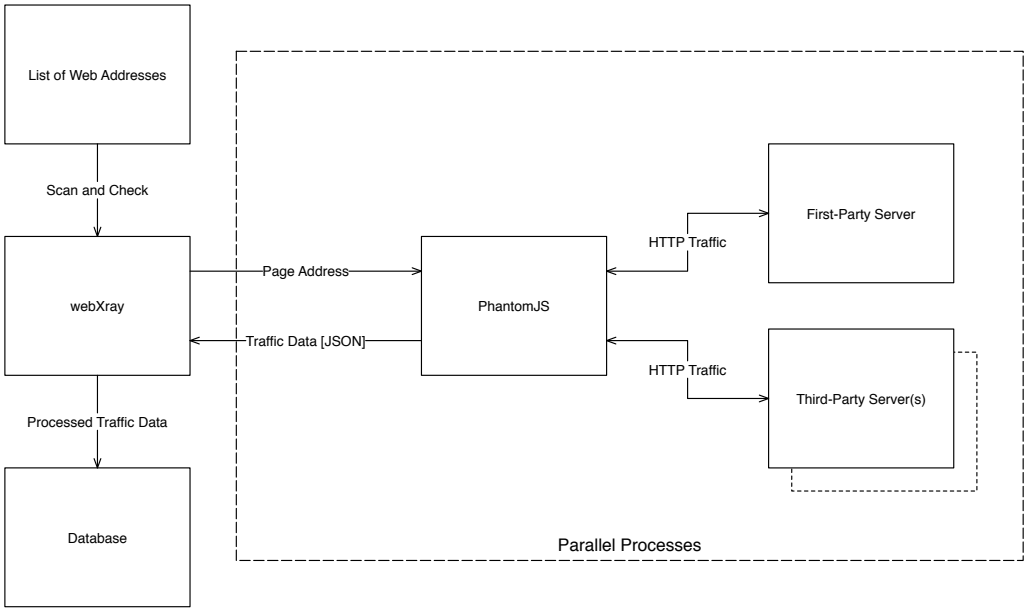

Figure 3.1: `webXray` *Architecture*

Figure 3 illustrates the basic workflow of a `webXray` collection sequence. First, a list of page addresses is passed to the program. This list of addresses is scanned

to ensure integrity and that they do not correspond to common binary files (e.g. Microsoft Office, PDF, etc). Next, page addresses are passed to PhantomJS using the Python "subprocess" module.

A small program manages loading the page address with PhantomJS. First, the program clears the cookie cache so that new cookies may be stored and analyzed. Second, it attempts to load the specified address, following redirects as necessary. In order to cope with pages delivered based on a specific user-agent string, the program randomly reports one of ten common user-agents to the page at load time. After possible redirects, the final address the browser is redirected to is recorded, this allows for the detection of sites which force browsers to use encrypted connections.

Due to the fact that many pages have a significant number of redirects and embedded elements which may be slow to download, the program waits for thirty seconds before closing the connection. Given that this waiting period makes it very difficult to analyze large volumes of sites in a reasonable amount of time, the Python "multiprocessing" module is used to instantiate many instances of PhantomJS in parallel. The number of instances scales linearly with available computer memory: doubling the memory results in twice the number of instances, and therefore doubles the speed.

As PhantomJS loads the page, HTTP network events are recorded using the "onResourceRequested" and "onResourceReceived" API calls. "onResourceRequested" indicates that the page has triggered a request to download an element. "onRe-

sourceReceived" provides information on if the requested data was returned, how long it took to download, and the size of the element measured in bytes.

One reason it is important to detect if a request returned data is that if content is being dropped at a network level it may indicate intentional blocking of the element. Measures of load time and element size are important in determining the impact of third-party requests on user experience in terms of time spent downloading a page as well as the costs of data transfer. This data is relied upon in the chapter which investigates news websites.

After PhantomJS has completed loading the page, data on requests, cookies, URL redirects, page title, and meta description are formatted as Javascript Object Notation (JSON)[15] data and passed back to the calling Python script for further processing. In cases where the page is unable to be loaded, errors are passed back and logged. The final step shown in the diagram is data being processed and stored in the database.

Collecting HTTP requests in bulk is a difficult process, and making sense of the data is no less complicated. Once data is returned from PhantomJS several steps must be taken. First, requested data must be analyzed to detect third-parties and stored. Second, mappings of third-party domains to their owners must be built.

When `webXray` processes the JSON data received from PhantomJS, cookies and element requests stored and linked to the pages which initiated their request.

[15]JSON is a means to store and transmit data, for details see: http://json.org

All cookies are stored as unique, and no further processing is done. Elements are evaluated as unique according to the full URL which they are downloaded from, and only the first instance of a given URL is stored in the database so that additional instances may reference the first. However, oftentimes a URL will be appear unique based on trailing arguments alone, most often this is the result of site identifiers being added to a common element URL.[16]

For example, the following unique URLs, "example.com/tracker.png?site=1234" and "example.com/tracker.png?site=5678", actually refer to the same file: "example.com/tracker.png". To cope with this, URLs are parsed to remove arguments and the element URL is stored separately for purposes of comparison. In order to facilitate in-depth analysis, file identifiers such as "png" and "js" are parsed so that that reports may be generated on the most popular images, Javascript files, and the like.

Cookies and elements are linked to the pages they correspond to in junction tables. These tables have a field for the page identifier, the cookie or element identifier, as well as a field which specifies if they are a result of a third-party requests. This system detects cases where an element may exist as a first-party on one site and a third-party on another. In order to determine what party classification is to be made, the domain of the page and the domain of the element are compared to see if they match, those which do not are third-parties and are marked appropriately.

[16]It is also a common technique to evade the browser cache by appending meaningless data to the end of a URL.

Parsing and comparing domains is not a trivial task. According to Mozilla, there is "no algorithmic method of finding the highest level at which a domain may be registered for a particular top-level domain [TLD] (the policies differ with each registry), the only method is to create a list".[17] Thankfully, Mozilla provides such a "Public Suffix List". This list is used by several major browsers to manage cookie settings (Firefox, Chrome, Opera, and Internet Explorer), and is also used by `webXray` to parse and compare domains between pages, cookies, and elements.

Once parsed, domains are stored in the "domain" table. This table includes fields for the Public Suffix so that pages and elements may be analyzed based on the country-code top-level domain (ccTLD), allowing for basic comparison of tracking between countries. This is an imperfect proxy however, as "com" is one of many popular TLDs used around the world.

Thus far, a system has been described for determining at a network level what information is flowing between two or more parties when loading webpages. Who these parties are, what they do with the data, and their motivations for collecting it (either intentionally or incidentally), are not revealed by a purely technical analysis. It is therefore necessary to determine what parties own given third-party domains, which is the core feature of `webXray`.

The first step in the process is to select which domains will be investigated. On small-sized analyses, such as 500 pages, the number of unique third-party domains

[17] https://publicsuffix.org/learn/

will be such that it is possible to investigate all of them. On larger analyses, of several hundred thousand or more pages, it is sufficient to select the 100 most frequently occurring domains for investigation. For any given analysis, the number of domains may be made fairly manageable.

Once domains are selected, they may be manually examined to determine ownership. The *whois* protocol specifies a means to "provide information services to Internet users", such as the ownership of a given domain.[18] The UNIX utility *whois* may be used to query the database in order to determine the party that has registered, and therefore owns, a given domain. For example, a whois query for "2mdn.net" reveals the "Registrant Organization" is "Google Inc." - thus, users who download elements from "2mdn.net" have their data sent to Google.

In cases where the organization name is unfamiliar (e.g. a niche advertising network) it is worth looking up further information on the company. This may reveal that the organization is actually a subsidiary of another company. It is also helpful to consult "Crunchbase", which is a database of technology companies and acquisitions. Crunchbase's motto is it allows one to "Discover innovative companies and the people behind them" - for the purposes of tracking data flows, it is certainly the case.[19]

The whois database does not always provide answers as it is possible hide ownership using an anonymizing service. In these cases it is sometimes sufficient to visit

[18]https://tools.ietf.org/html/rfc3912

[19]See: https://www.crunchbase.com/

a given domain in a web browser and be directed to the company which owns the domain. Yet, sometimes the domain in question will not have a web page. In those cases one may ping the domain to find the IP address, and then attempt to find the owner of that address.[20] It is also possible to find information using in-depth detective work; in one case the owner of a domain was revealed by reading obscure API documentation which revealed the address in question.

As many companies receiving data go to great lengths to keep their activities hidden, it is a laborious process to track down all of the entities who receive data. In order to reduce duplicated effort, domain ownership data is stored in a small JSON-formatted database and is part of the core `webXray` distribution.

Once data flows and domain ownership has been established, `webXray` generates a variety of reports which reveal the companies which receive data, the most frequent types of third-party elements, how many companies and elements users are exposed to on average, as well as variations between sites hosted on different top level domains. These reports may then be extended to further explore the implications of findings.

[20] 'Ping' is a standard UNIX utility which sends a network request to a given address to read a reply which indicates the machine is available (machines can also ignore 'pings'). By sending a 'ping' to a given domain, the IP address of the reply may be discovered in the response.

Extensibility, policyXray, & Malware Analysis

One of the strengths of the webXray data model is it allows for the addition of new types of data which may be easily joined to the page, element, or domain tables. The basic process for this is to export the relevant data needed to query outside data sources, gather the new data, add a table to the webXray database for the specific type of data and use a Python program to insert the new records with the appropriate record ids. Once this is performed, customized reports may be generated using the new data sources. This capability has been employed to audit privacy policies and detect the presence of known malware-associated domain names with pages.

The webXray data model is a relational database schema designed to allow for efficient storage as well as easy extensibility. The basic approach is to segregate information on pages, cookies, and elements into their own tables. The element and cookie tables each have a boolean field "is_3p" which allows queries to easily discriminate between first- and third-party requests. Likewise, the "domain" table segregates information so it may be joined to the "org" table where domain ownership is determined. The domain table may be linked to ancillary programs to query outside datasources in order to provide further depth to findings. Figure 3 lays out the core schema in detail.

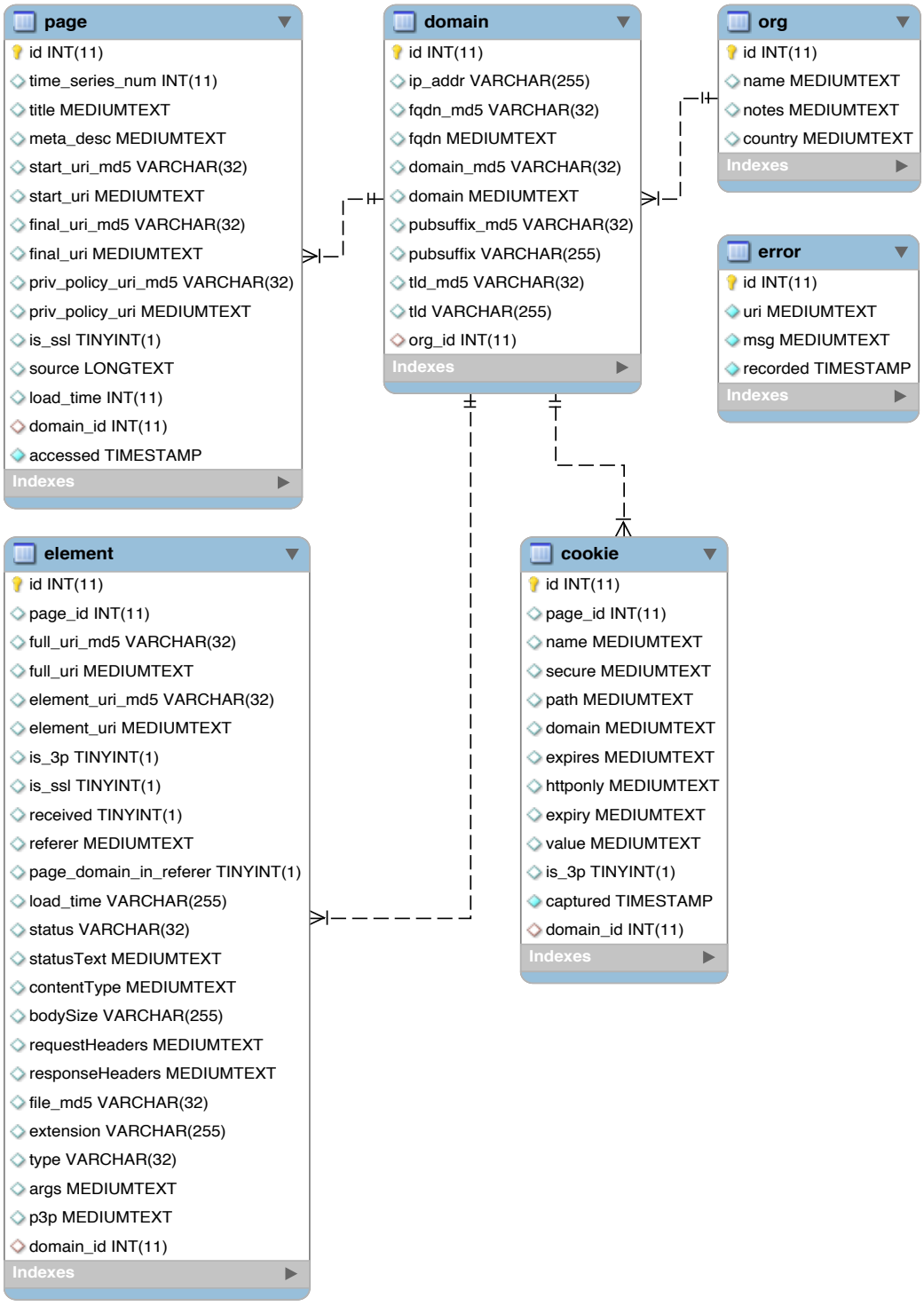

Figure 3.2: *webXray* Data Model

A second module for `webXray`, `policyXray`, highlights the value of the relational approach. The purpose of `policyXray` is to download, process, and analyze privacy policies to determine if the companies detected by `webXray` are mentioned in the policy. This is performed by extracting the policy text and determining if the companies receiving user data on the page are mentioned in the policy. It is a deceptively simple approach which requires a complex implementation.

When `webXray` is monitoring HTTP request traffic, it also analyzes the links on a page in order to find the associated privacy policy. This is performed by finding the first link with the English-language text "privacy policy", if this fails, the first link with the text "privacy" is stored. `webXray` does not analyze privacy policies at the time of collection due to increased processing and network overhead, instead `policyXray` may be run upon completion as a separate process. Like `webXray`, `policyXray` is designed to run many instances in parallel, facilitating large-scale collection.

While downloading policy pages is fairly straight-forward, policy text must be isolated from the larger HTML document before it may be analyzed. The reason it is insufficient to search the full HTML document is illustrated when looking for a mention of a social media company such as Facebook. It is possible that a mention of "Facebook" in the HTML is in reference to a "Like us on Facebook" link, rather than a specific policy disclosure.

Extracting article content from a web page is a known difficult problem in the

field of natural language processing, however the "Readability.js" Javascript library provides an automated method for weighting various page elements to determine what parts of the page are article content (e.g. by excluding sections of the page likely to contain navigational elements). The Safari and Firefox browsers both use this library, and Mozilla hosts the library on GitHub.[21] For `policyXray`, a Python port of Readability.js is used to extract policy content from the page.

Once `policyXray` has downloaded the HTML document and extracted the policy text, a sanity check is run to determine if the text appears to be a privacy policy. This is done by checking for the terms "privacy" and "cookie" in both the policy text and the page title, if these terms are both missing, the policy is discarded. In cases where Readability.js is unable to parse the page, it is also discarded.

In order to verify if detected data flows are disclosed in a given privacy policy, `policyXray` searches the text to find either the name of the parent company which has been identified by `webXray` or, if the company has subsidiaries, if the subsidiaries are mentioned. Due to the complexity of ownership patterns and business acquisitions in the OBA sector, this is a non-trivial undertaking. After running this final process, `policyXray` is able to determine the percentage of data flows disclosed on a per-page as well as population-level basis. Other analyses may be be performed as well.

Given that `webXray` is able to produce reports of which advertising network

[21] https://github.com/mozilla/readability

domains are connected to a given population of pages, it is possible to take domains and determine if they have been recently used for delivering malware. While there are numerous normative objections to surveillance-based advertising, advertisers may make reasonable arguments they contribute to online commerce. However, there is no argument which may be made in favor of facilitating cybercrime. Thus, using `webXray` data to reveal malware exposure provides a strong argument that self-regulation is failing. Furthermore, any data flows resulting in criminals gaining access to users' computers is a *de facto* violation of privacy.

To determine the presence of malware within a `webXray` dataset, data from VirusTotal is used. The company is a Google subsidiary and provides a "free online service that analyzes files and URLs enabling the identification of viruses, worms, trojans and other kinds of malicious content detected by antivirus engines and website scanners" using aggregated data from "different antivirus engines, website scanners, file and URL analysis tools and user contributions".[22] The company provides free public API access to this service (though with a very slow 4 requests/minute limit).[23] The public API provides access to a number of reports, the one most useful for enhancing `webXray` data is the domain report API which aggregates data from over 60 anti-virus databases.

The VirusTotal domain report returns several fields of data, the most important of which is "response_code" which specifies if a given domain is in the VirusTotal

[22]https://www.virustotal.com/en/about/

[23]https://www.virustotal.com/en/documentation/public-api/

database. For any domains in the database, a collection of URLs which have been scanned is returned in the "detected_urls" field. For each URL in the collection the number of scans across anti-virus services is given along with the number of positive matches for malware. This makes it possible to say if a given domain has been associated with malware by any anti-virus report. However, a single match may be the result of a mistake by a given source, thus it is also necessary to calculate the number of URLs which have been identified as having malware by two or more sources, which greatly reduces the likelihood a domain was associated with malware in error. Finally, the API has a field "categories" which ranges from innocuous classifications such as "educational institutions" to alarming titles such as "known infection source". While VirusTotal is a good resource to measure malware exposure on a given page population, it alone does not reveal if a given population is worse than average.

Indeed, when examining a given page population, it is helpful to know if third-party data flows within the specified context are in some way "normal", or if they constitute a particular aberration. However, it is difficult to define what is the "normal" amount of data flows on the web as there is no canonical source which reveals the average amount of tracking on all websites. Thus, to gain such measures, a population of commonly visited websites must be constructed and analyzed.

As noted above, the Alexa company provides a list of the one million most popular websites globally; this list is the best available means to determine what a

"normal" amount of tracking across highly traffic sites. However, the list is highly variable as a "long tail" of websites changes quite frequently and web masters are constantly trying to improve their global rank through questionable means. Despite this, the top 100,000 sites in the list are more stable, and form a better population from which to draw global averages.

Speed, Cost, and Limitations

The `webXray` method has many strengths in terms of speed and cost, as well as several limitations. A key consideration in the design of `webXray` was to maximize efficiency in order that cheap "cloud" based virtual machines could be used to perform analyses. Based on current prices for cloud servers, one may analyze the top one million websites identified by Alexa in 30 hours for roughly $32, which is a fraction of the time and cost of using similar tools (Englehardt and Narayanan, 2016).[24] The benefit of this approach is that companies which expend billions of dollars to covertly track users across the web may be revealed for a comparatively minuscule sum.

`webXray` has numerous limitations. It should first be stated that the main limitations all have the outcome of potentially *undercounting* the amount of data flows taking place - but in no case should a data flow which is not present be recorded. That said, data flows can be missed for a variety of reasons. First, given

[24] Pricing estimate generated on May 31, 2016.

the rapid rate of website ingestion, it is entirely possible that the IP address being used in a given analysis would be blacklisted[25] - in this case the number of requests would go down as a result of Javascript or iFrame elements which spawn additional requests not loading.[26] However, some indication of this would be found in lower rates of requests being received.

Second, PhantomJS does not support a variety of browser plug-ins such as Adobe Flash, Microsoft Silverlight, or Java Applets. Given that none of these plug-ins are supported on popular mobile browsers, their market share is rapidly decreasing: nonetheless it would be preferable to capture them. The trade-off in this case is utilizing a full browser with plug-in support adds significant cost, complexity, and time.

Finally, advertising companies expend tremendous effort and resources on making their tracking mechanisms difficult to detect and circumvent; thus, there is a high probability the methods used by `webXray` simply are not sufficient to cope with intentional blocking designed to hide various tracking activities from analysts. Another less malevolent reason for advertisers to block automated browsers is to cope with "click fraud", a technique by which criminals manipulate ad networks to

[25] "Blacklisted" in this context meaning servers getting requests from the specified IP address would ignore them.

[26] iFrames are a means by which one HTML page may be inserted into another, thus a single browser window may actually be displaying content from several HTML files, all of which may be initiating additional HTTP requests.

gain unearned payment. Indeed, there exists some suggestion that PhantomJS may be intentionally blocked by some ad networks (Englehardt and Narayanan, 2016); however, user-agent randomization appears to alleviate much of this blocking.

Additional limitations are to be found in measures of time and data size. Measures of how long elements take to download are dependent on the speed of the network connection, congestion, and other factors. Thus, in no way should the time taken to download an element be taken to have any absolute meaning - it may only be used as a relative measure between two or more network events during the same analysis - and in that case only with the caveat that comparisons are roughly approximate. Second, given the manner in which PhantomJS loads elements and reports sizes, it is possible there is some skew in the actual size of the elements, thus the measures should be used as approximations.

Measures of page load times are complicated by the fact that many pages continue to initiate requests for additional content in the background as the user scrolls down a page - thus, when a given page is "finished" loading is a somewhat imprecise concept, and one that is poorly reflected by the available measure which is based upon the time delta between the first request event and the final receive event. Also, given the thirty second timeout, pages that take longer than this time are not accurately measured.

Finally, `policyXray` is severely limited by the fact that only English-language links to privacy policies may be followed. On large page populations which include

many languages, this results in incomplete coverage. Furthermore, while Readability.js performs well, it is not always able to cope with the structure of a given page and is unable to work flawlessly. Furthermore, in some cases pages have been found which encode text in a manner which makes it exceedingly difficult to extract and these pages are often discarded by the sanity check.

Conclusion

This chapter has described newly-developed methodologies used to construct page populations which reflect social contexts, investigate pages for data flows which violate social norms, and auditing the policies of web sites and potential malware exposure. These methodologies facilitate pursuing the answers to the complex social questions asked in the following chapters.

Chapter 4

On the Impossibility of Accepting the Unknown: A Web-Scale Analysis of the Failure of Notice and Choice

"The [Network Advertising Initiative] Code of Conduct, is a set of self-regulatory principles which all NAI member companies must agree to uphold in order to be members. The Code requires notice and choice with respect to Interest-Based Advertising, limits the types of data that member companies can use for advertising purposes, and imposes a host of substantive restrictions on member companies' collection, use, and

transfer of data used for Interest-Based Advertising."

Network Advertising Initiative, Code and Enforcement[1]

Advertisers collect information about users in order to show users advertisements more "tailored" to their interests using the techniques of Online Behavioral Advertising (OBA). In exchange, advertising revenue subsidizes the vast volumes of media content which users expect to receive for free. The above quotation from the industry group Network Advertising Initiative suggests that the exchange of personal behavioral data for media content is premised on an implicit agreement referred to as "notice and choice". In this model, users are notified that personal data is used for advertising and are given a choice to halt such use (often known as an "opt-out"). The common industry picture is that users are knowingly, and happily, giving up personal information in exchange for information goods.

While OBA is fairly new phenomena, the concept of notice and choice is drawn from a much older source. The Fair Information Practice Principles (FIPPs) are an established set of guidelines for the processing of personal information by computerized systems. The FIPPs are over forty years old and the underlying principles have been adapted and modified in regulations around the globe. Nissenbuam has observed that the original FIPPs functioned "like a bill of rights" which was designed to protect "private individuals (data subjects) against large government and private sector institutional actors" (Nissenbaum, 2004).

[1]http://www.networkadvertising.org/code-enforcement

FIPPs have evolved over time, starting with five broad principles, and subsequently growing to as many as nine in later adaptations. The OBA industry, however, has chosen just two FIPPs-like principles, "notice and choice", to form the primary basis of their self-regulatory regimes. This has drawn wide rebuke from privacy scholars and advocates who feel that industry self-regulatory efforts neither embody the normative spirit of FIPPs, nor protect privacy in any meaningful way. The literature is full of historical, legal, and philosophical objections to the "notice and choice" paradigm.

Despite the large volume of academic literature on the topic, surprisingly little has been done to identify the actual companies which receive data through online tracking and verifying that the vaunted "notice" component of self-regulation is in fact occurring as promised. Likewise, while much has been written about efforts such as the "Do Not Track" standard for signaling a user's "choice" in data preferences, there is little information available on the degree to which websites and advertisers follow or reject the standard. Overall, disputes over legal minutiae have kept focus on the question if "notice and choice" is acceptable, rather than if it is being practiced in a way which respects even the most basic formulations of meaningful norms.

Some of the blame for this situation is an overabundance of versions of FIPPs, and complex language which has led to confusion over what exactly constitutes "notice", let alone "choice". For this study, a common sense approach has been

taken and simple questions are asked. First, and foremost, who are the companies receiving user data and are they known to most users? Second, if a diligent user were to read the privacy policy of a given website, would she or he be able to understand the text of the policy and learn the names of the companies receiving data? Third, do the companies receiving user data commit to respect the "choice" expressed by the "Do Not Track" standard? Finally, do the various companies collecting data utilize basic security practices to protect user data from malicious parties? Only by investigating these questions directly will we discover if "notice and choice" is actually being practiced today.

While the questions posed above are fairly simple, the process for answering them requires significant technical effort and custom tools were developed for this study. These tools, and the data produced by them, represent a significant methodological advancement in the field of online privacy research. In order to provide clarity on the dominant practices occurring on the web today, this study is rooted in a large-scale analysis of data flows on one million web sites with a focus on attributing data flows to the specific companies which receive user data. In order to get to the core of how notice and choice is or *is not* being practiced, the largest ever assembled corpus of privacy policies, those corresponding to over 200,00 websites, has been collected and analyzed. This analysis is the first of its kind conducted at this scale. Network traffic has further been inspected to determine if minimum security precautions are being followed. Finally, rather than treating the industry as an

undifferentiated whole, focus has been applied to the practices of the 25 companies which receive the bulk of user data in order to reveal variations in practices.

The underlying normative implication of this study is that if the OBA industry's crippled version of the FIPPs guidelines, "notice and choice", is not functioning in practice, then the broader normative goals embodied in FIPPs fail as well. Ultimately, the FIPPs goal of placing controls on the behavior of powerful actors to preserve individual rights is at stake.

This chapter begins with an accounting of the origin and history of FIPPs as a normative grounding for evaluating data practices. Next, the process by which industry has manipulated the ideas behind FIPPs to generate self-regulatory frameworks is explained. This is followed by a summary of the dominant objections to the "notice and choice" formulation. In order to facilitate a nuanced evaluation, research questions are refined and population designs are established.

The evaluation demonstrates with a high degree of certainty that "notice and choice" fails in practice as well as in theory. The industry simply does not regulate itself in a way which meets the normative goals of FIPPs. These underlying failures set the stage for subsequent chapters which will further explore how OBA has had deleterious impacts on the sanctity of personal medical information and the role of the press in a democratic society.

Normative and Regulatory Background

There is an oft-made claim made by those in the business of monetizing personal data that privacy is a somehow modern invention, and one with a limited lifespan at that. Such a viewpoint is not only ignorant of ancient history where privacy was valued by the Greeks and Egyptians, but it is also ignorant of *recent* history. Decades before the first tracking pixel was a glean in the eye of a marketer, regulators and theorists realized that the proliferation of computerized databases would have profound impacts upon society. They sought to advance concepts and policies to protect individuals; however, such efforts have been in constant struggle against those who seek to profit from the erosion of privacy rights.

In 1973 the U.S. Department of Health, Education, and Welfare convened a committee on "Automated Personal Data Systems" to study the impacts of the "growing use of automated data systems containing information about individuals" by both "public and private sector organizations" (Gellman, 2016). The committee subsequently produced the report "Records, Computers and the Rights of Citizens" which included a set of privacy principles called the *Code of Fair Information Practices* (Gellman, 2016). This code contained five guidelines which came to be known as the Fair Information Practice Principles, or simply "FIPPs". The code is reproduced below:

HEW Code of Fair Information Practices

1. There must be no personal-data record-keeping systems whose very existence is

secret.

2. There must be a way for an individual to find out what information about him is in a record and how it is used.

3. There must be a way for an individual to prevent information about him obtained for one purpose from being used or made available for other purposes without his consent.

4. There must be a way for an individual to correct or amend a record of identifiable information about himself.

5. Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure the reliability of the data for their intended use and must take reasonable precautions to prevent misuse of the data.

Aside from the use of the male pronoun, the original FIPPs were incredibly forward-looking and anticipated many of the social issues raised by the era of "big data". Privacy advocate Marc Rotenberg has observed that "more broadly, the Fair Information Practices set out an approach to the design of information systems that ***embeds certain normative political views***" (emphasis added) (Rotenberg, 2001). Legal scholar Fred Cate has written that the FIPPs "reflected a wide consensus about the need for broad standards to facilitate both individual privacy and the promise of information flows in an increasingly technology-dependent, global society" (Cate, 2006).

The HEW report led directly to the 1974 U.S. Privacy Act which applied the FIPPs to the federal government. However, the Privacy Act is but a footnote in the longer history of the FIPPs which have evolved, expanded, and been applied in a number of regulatory and cultural contexts.

In 1980, the original five FIPPs were extended to eight by the Organization for Economic Co-operation and Development in the document *OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data*. As an international organization, the OECD guidelines aimed to influence the development of national privacy laws in countries which did not yet have them. According to Cate, "in this aspiration they have undoubtedly succeeded because most of the dozens of national and regional privacy regimes adopted after 1980 claim to reflect the OECD Guidelines" (Cate, 2006). In many ways the OCED Guidelines became the new "benchmark for assessing privacy policy and legislation" and included eight principles which expanded on the original five: "Collection Limitation; Data Quality; Purpose Specification; Use Limitation; Security Safeguards; Openness; Individual Participation; and Accountability" (Rotenberg, 2001).

Throughout the intervening decades, FIPPs and OECD-influenced regulations spread widely. The 1995 EU Privacy Directive relied on the principles and "ensured the spread of [FIPPs] throughout Europe" (Gellman, 2016). Canada implemented a national standard for data protection in 1995 which likewise reflected "the international consensus on [FIPPs]" (Gellman, 2016). In 2000, the U.S. Department

of Health and Human Services "relied upon [FIPPs]" in crafting the medical privacy rules mandated by the Health Insurance Portability and Accountability Act (HIPAA) (Gellman, 2016).[2] When the U.S. Department of Homeland Security was established in 2002, a privacy office was created and given the "responsibility for assuring compliance with fair information practices" (Gellman, 2016). The Asia-Pacific Economic Cooperation (APEC) forum expanded the concept of FIPPs to include nine principles in 2004 as a "conscious effort to build on the OECD Guidelines" (Cate, 2006).

It is no stretch to assert that "modern data protection law is built on 'fair information practice principles'" (Cate, 2006). Likewise, it is clear that they represent not only the political norms of the 1970s United States, but norms which have received adoption and support around the globe. However, these norms, and the regulations based on them, have themselves been unable to stop the large-scale erosion of privacy brought about by Online Behavioral Advertising (OBA). The reason for this is that both U.S. regulators and the OBA industry have promoted significantly weakened interpretations of FIPPs in the online space.

As noted in the introductory chapter, the emergence of Online Behavioral Advertising in the mid- to late-1990s represented a foundational shift in the business and practices surrounding media and advertising. These shifts were directly tied to the accumulation of personal web browsing data in electronic databases, precisely

[2]HIPAA is covered further in the next chapter.

the type of activities which FIPPs and related frameworks anticipated. However, due to a dilution of FIPPs principles, effective regulation of OBA has proven elusive.

While the original conception of FIPPs included five guidelines, the OECD eight, and the APEC nine, a 2009 U.S. Federal Trade Commission report titled "Self-Regulatory Principles For Online Behavioral Advertising" includes only four. This report was not the first or last word from the FTC on the topic of FIPPs. According to Gellman, "from 1998 through 2010, the Commission's description of FIP[P]s has been consistently inconsistent" (Gellman, 2016). However, the 2009 version may be the most influential as it is the version which has been favored by industry as a basis for their own self-regulatory guidelines.

The "Self-Regulatory Principles for Online Behavioral Advertising" is a document jointly developed by the trade groups of the American Association of Advertising Agencies, Association of National Advertisers, Council of Better Business Bureaus, Direct Marketing Association, and the Interactive Advertising Bureau. The document espouses several principles and claims to "correspond with the 'Self-Regulatory Principles for Online Behavioral Advertising' proposed by the Federal Trade Commission in February 2009". The 2009 FTC guidelines favored by industry are as follows:

FTC 2009 Self-Regulatory Principles for Online Behavioral Advertising

1. **Transparency and Consumer Control**: Every website where data is collected for behavioral advertising should provide a clear, concise, consumer-friendly, and

prominent statement that (1) data about consumers' activities online is being collected at the site for use in providing advertising about products and services tailored to individual consumers' interests, and (2) consumers can choose whether or not to have their information collected for such purpose. The website should also provide consumers with a clear, easy-to-use, and accessible method for exercising this option. Where the data collection occurs outside the traditional website context, companies should develop alternative methods of disclosure and consumer choice that meet the standards described above (i.e., clear, prominent, easy-to-use, etc.)

2. **Reasonable Security, and Limited Data Retention, for Consumer Data**: Any company that collects and/or stores consumer data for behavioral advertising should provide reasonable security for that data. Consistent with data security laws and the FTC's data security enforcement actions, such protections should be based on the sensitivity of the data, the nature of a company's business operations, the types of risks a company faces, and the reasonable protections available to a company. Companies should also retain data only as long as is necessary to fulfill a legitimate business or law enforcement need.

3. **Affirmative Express Consent for Material Changes to Existing Privacy Promises**: As the FTC has made clear in its enforcement and outreach efforts, a company must keep any promises that it makes with respect to how it will handle or protect consumer data, even if it decides to change its policies at a later date. Therefore, before a company can use previously collected data in a manner ma-

terially different from promises the company made when it collected the data, it should obtain affirmative express consent from affected consumers. This principle would apply in a corporate merger situation to the extent that the merger creates material changes in the way the companies collect, use, and share data.

4. **Affirmative Express Consent to (or Prohibition Against) Using Sensitive Data for Behavioral Advertising**: Companies should collect sensitive data for behavioral advertising only after they obtain affirmative express consent from the consumer to receive such advertising.

A cursory review reveals why industry has embraced the FTC 2009 guidelines. The first principle calls for a "consumer-friendly", but otherwise poorly defined, notice that data will be used for advertising. This principle further specifies that "consumers" should be able to "choose whether or not to have their information collected". Neither aspect of this principle calls for specifics on how the processes should work in practice. The second principle of security and data retention lacks specifics of the former, and in regards to the latter allows that data may be held for "as long as is necessary to fulfill a legitimate business...need". Given that OBA is premised on building behavioral profiles of users, "necessary" may reasonably mean "indefinite" under the principle. The third principle merely states that advertisers shouldn't change their terms retroactively, which is less an appeal to protect essential human dignity as it is a means of stating the obvious nature of a contract.

The fourth principle is the shortest and most vague, yet it hints at recognition

of essential privacy rights insomuch as it provides acknowledgement that certain types of data may be considered "sensitive". However, it fails to place specific types of data off-limits, and to the degree they exist, it merely suggests advertisers ask before they use them. If this were not lax enough, the industry has interpreted that guideline as an incredibly narrow set of data applying to information about children (already regulated by federal statue[3]) and "financial account numbers, Social Security numbers, pharmaceutical prescriptions, or medical records".

The result of the FTC's weakened interpretations of FIPPs is that a comprehensive set of normative principles have been essentially reduced to two pro-industry suggestions: that consumers receive notification of data transfers (primarily in the form of privacy policies), and that they be given an ill-defined choice in what happens to their data. This forms the basis of the "notice and choice" paradigm and reflects a decision by the FTC to embrace "a relatively recent creation of the U.S. marketing industry" which is at odds with internationally accepted versions of FIPPs (Rotenberg, 2001).

In essence, the FTC has chosen not to regulate privacy practices, but instead "focused virtually all of its...efforts on getting websites to post privacy policies and its enforcement efforts on suing website operators when they fail to follow those policies" (Cate, 2006). This approach has been met with widespread criticism from privacy scholars and advocates.

[3]Specifically, the Children's Online Privacy Protection Act of 1998

There are numerous objections to the notice and choice paradigm which mainly have to do with the the fact that the system does not respect the essential privacy rights envisioned by FIPPs, the limits of what users consent to are often unbounded and conceptually flawed, and the notices are difficult to understand and therefore impossible to consent to.

While FIPPs was originally conceived as a normative statement on privacy rights, the notice and choice interpretation of FIPPs substitutes "maximizing consumer choice for the original goal of protecting privacy while permitting data flows" (Cate, 2006). The focus on "consumer choice" has resulted in "the energy of data processors, legislators, and enforcement authorities [being] squandered on notices and often meaningless consent opportunities, rather than on enhancing privacy" (Cate, 2006). Furthermore, when compared to other areas of health and safety where minimum standards are enforced, the notice and choice approach transfers "the protection of privacy from the legal realm, and from an emphasis on the articulation of rights and responsibilities, to the marketplace, where consumers would now be forced to pay for what the law could otherwise provide" (Rotenberg, 2001).

Beyond the erosion of rights in favor of market mechanisms, the notion of consent is problematic as "consent legitimizes nearly any form of collection, use, or disclosure of personal data" and places few, if any, limits on how companies may collect or manipulate data, thereby denying users the type of "meaningful control" which consent implies (Solove, 2012). Furthermore, it is difficult to understand the

increasingly complex practices detailed in policies. Barocas and Nissenbuam have observed that "users who are subject to [online tracking] confront not only significant hurdles but full-on barriers to achieving meaningful understanding of the practice and uses to which they are expected to be able to consent" (Barocas and Nissenbaum, 2009). Likewise, given the proliferation of complex machine learning techniques, Barocas and Nissenbaum conclude that in the era of big data "when analysts can draw rules from the data of a small cohort of consenting individuals that generalize to an entire population, consent loses its practical import" (Barocas and Nissenbaum, 2014).

Even assuming that notices offer bounded and technically comprehensible limits on data processing, the application of common readability metrics such as Flesch-Kincaid reveals that they remain difficult to read. A study from 2002 looked specifically at privacy policies on health websites and determined that "of the 80 Internet health Web sites studied, 30% (including 23% of the commercial Web sites) had no privacy policy posted. The average readability level of the remaining sites required 2 years of college level education to comprehend, and no Web site had a privacy policy that was comprehensible by most English-speaking individuals in the United States" (Graber et al., 2002). Likewise a more general longitudinal study of privacy policies in the years 2001 and 2003 found that "based on the Flesch–Kincaid grade level, the average grade level for 2001 was 11.2...in 2003, the mean grade level increased to 12.3...finally, 54.8% of the privacy notices reflected an increase in grade

level by .5 from 2001 to 2003." (Milne et al., 2006). These trends continued to a 2009 study which found that the Flesch-Kincaid Reading Ease score for six privacy policies all fell within the "difficult" range (McDonald et al., 2009).

The impact of the difficulty of reading privacy policies is that it directly undercuts the ability of users to consent to data practices. Reidenberg et al have argued that "privacy policies may be misleading the general public and that those policies could be considered legally unfair and deceptive. And, where websites are not effectively conveying privacy policies to consumers in a way that a 'reasonable person' could, in fact, understand the policies, 'notice and choice' fails as a framework" (Reidenberg et al., 2014). Likewise, Blanke has argued that "just because someone clicks a button, it does not necessarily follow that there is even an objective intent to enter into a contractual relationship...courts should not find genuine assent merely because (objectively) a button was clicked" (Blanke, 2006).[4] Finally, privacy policies are often hard to find, which means that "if the notice is never received by the consumer, the choice it provides is meaningless."(Cate, 2006)

Overall, critics contend that the notice and choice paradigm "falls well short of" the standards advocated by FIPPs (Gellman, 2016). However, due to the proliferation of many versions of FIPPs over the past forty years, and the vast complexity of the online tracking ecosystem, there is little consensus as to the exact nature and

[4]Blanke's argument was specifically in regards to spyware which was downloaded and run on a user's computer, the difference with web tracking is primarily an issue of computer architecture rather than consent-driven.

mechanisms by which notice and choice fails.

Research Design and Objectives

In absence of a single version of FIPPs, let alone agreement between and among industry, regulators, and advocates of what privacy protections in the Online Behavioral Advertising (OBA) space should look like, there is no accepted metric by which to judge the privacy practices of either websites or advertising networks. The closest to consensus is that the OBA industry and US regulators broadly endorse the "notice and choice" paradigm. However, the definitions and caveats involved in "notice and choice" render it particularly difficult to use in an empirical manner. In terms of basing research objectives on specific existing frameworks or policies there simply is nothing tangible enough to use as a measurement device.

In absence of accepted metrics, it is possible to move beyond legal theorizing and use a common sense approach. This study grounds fairly straight-forward questions regarding notice, choice, and security in a large-scale empirical study of 90 million records of data flows drawn from one million websites as well as over 180,000 privacy policies which correspond to over 200,000 websites. This represents the largest study of its kind, and proceeds in several steps.

The first step in this study is to provide an audit of the current state of personal data flows on the web. This provides a basis from which to gauge the spread and nature of OBA on the web, as well as identify the companies which are the primary

recipients of data. For this the `webXray` software platform is used to scan the top one million Alexa websites. `webXray` identifies data flows which transmit the address of the current page a user is viewing to any of 183 identified companies, the top 25 of which have been analyzed in depth in their observed practices as well as the content of their privacy policies.

Once the data pool has been established, the issue of notice comes to the fore. There are a number of questions in this area. First, it is entirely possible that the companies collecting user data are ones which users may have pre-existing consumer relationships and may have already agreed to the terms of their privacy policies. In that case, "notice" may pre-exist, and users may be aware that data is collected from external websites. Thus, the first metric is to evaluate of the companies receiving data, how many have consumer, as opposed to business, services and products?

Given the set of companies examined is likely to include many without consumer services, the next question is if the user can learn a given company is receiving data by examining the features of the website they are visiting to find indications of data flows. One possible indication is the industry-sponsored "AdChoices" icon which is a small blue triangle sitting in the corner of a targeted advertisement. The icon is designed to link the user to the entity which placed the advertisement along with some explanation of how it was targeted. However, researchers have found that "the purpose of these icons, to provide information to consumers, eluded participants, even when the icons were shown in context on an advertisement" (Ur et al., 2012).

Relying on the "AdChoices Icon" is thus a dead end.

The next possible way a user could expect to learn of the companies which receive user data is reading the privacy policy of a given website. Indeed, this is the most straight forward approach possible. Locating and reading one million privacy policies, and searching them for the names of companies is a task which a person, or a team of people, would find impossible. Thus, findings on company-directed data flows found using `webXray` are used as the basis for a `policyXray` analysis of policies relating to over 200,000 sites to determine if companies receiving data are mentioned. This approach provides both a population-level view of trends and practices in site policies, as well as per-company data which reveals the varying degrees to which companies are represented in policy texts.

A key logical component of the "notice and choice" paradigm is that notifications take a form which facilitates users' understanding of available options and attendant choices. Again, following a common sense approach, it is clear that notices must not represent an undue time burden to read and may only facilitate choice if they are easy to read for an average person. Thus, length and readability metrics for the policies of the top 25 companies receiving user data are evaluated to determine if the policies can be read in a reasonable amount of time and understood by an average reader, thereby facilitating notice.

There is significant debate in what constitutes "choice" in the OBA context. The industry viewpoint is that users have a choice to not see targeted advertisements, but

do not have a choice about data collection (though they do not state it so plainly). Critics roundly reject this formulation, and even the FTC has advanced the idea of the "Do Not Track" mechanism for users to signal to data collectors that they wish for their data not to be collected. "Do Not Track" is a feature in all modern web browsers which sends a message to data collectors that a user is signaling a choice in data collection practices. It is up to the companies to either respect or ignore this signal. Thus, in order to verify if an accepted choice mechanism is being respected, both the full set of policies as well as those from the top data collectors are evaluated to see if they mention and respect the "Do Not Track" mechanism.

Lastly, while not a component of "notice and choice", nearly all versions of FIPPs, as well as industry self-regulations, indicate that data security is a core value. Therefore, all data transfers are evaluated to determine if user data is protected from malicious parties using commonly deployed Secure Socket Layer (SSL) encryption.

Taken together, the various components of this methodology shed light on the dominant data collection practices on the web today, the companies which receive data, the degree to which users are reasonably notified, able to make choices about data collection, and can be confident their data is secured against interception.

Research Findings

Top-Level Audit of Information Flows

The home pages of the most popular one million websites as determined by Alexa were analyzed in January, 2017. 95% of sites were successfully analyzed and 5% failed to load. This failure may be attributed to the top sites lists containing sites which were not available at the time of analysis or had taken measures to defeat automated analysis. The pages analyzed yielded a set of 16,809,296 cookies, 11,974,851 (71%) of which were set by third-parties and represent a likely vector for user tracking and ad delivery. 91,321,943 elements were requested, of these 54% were requested from third-party domains meaning that over half of all network traffic potentially exposes user behavior to unknown third parties. Finally, 91.17% of all sites analyzed initiated requests to external parties, which indicates that users should view the loss of personal privacy as a routine occurrence when browsing the web.

Additionally, user data is often transferred to many parties in tandem. Further analysis determined that sites which *do* make requests to third parties contact 12.8 distinct domains on average - indicating that not only is user data being frequently shared, but it is often being shared with many parties simultaneously. 70.33% of sites spawn third-party cookies which may be used to track users using traditional methods, and 86.69% of sites include third-party Javascript which may represent the

increasing trend towards fingerprinting and other advanced techniques. It should be noted that since this study used a specified population rather than a statistical sample, the confidence for these numbers is 100%, but these results *only* represent the Alexa top one million - *not* the entire web.

In the same way that FIPPs have been adopted around the world, third-party web tracking is also a global phenomena. In addition to the top level trends, sub-analyses were conducted on the 10 most common top-level domains (TLDs) in order to determine how wide-spread tracking was around the world. The top TLDs include the well-known com, net, and org as well as several country-code top-level domains (ccTLDs) such as Russia (ru), Germany (de), Japan (jp), United Kingdom (uk), Brazil (br), India (in), and Iran (ir). Sites ending in a ccTLD have been registered within a given country - however, popular sites within a given country often include com and org sites, so this is *not* a perfect proxy for what a user from a given country is exposed to (i.e. do not assume Russians only visit *ru* sites).

When comparing the amount of third-party data flows by country, there was not much range in the percentage of sites which had at least one third-party element: Russian sites had the most with 96.26%, and Iranian sites had the least with 82.7%. However, in regards to the average number of domains contacted, there was a large spread with Brazilian sites contacting 15.28 on average and Iranian sites less than half, 6.25. There was also a wide disparity in the percentage of sites spawning third-party cookies: Russia was again in first place with 92.7% of sites spawning

cookies, whereas far fewer Iranian sites did so (64.26%). Finally, there was also a significant difference in the percentage of sites with third-party Javascript: at the high end were Russian sites with 92.3%, and at the low end were Iranian sites with 64.26%. These findings are presented in greater detail in Table 1.

Table 4.1: Findings Summary

TLD	N	% Total	% W/3PE	% W/Cookie	% W/JS	Ave. Domains Contacted
*	947356	100	91.17	70.33	86.69	12.8
com	468935	49.5	92.48	72.17	88.31	13.78
net	51193	5.4	87.77	67.67	82.11	12.84
ru	48042	5.07	96.26	92.7	92.3	12.1
org	44240	4.67	89.18	62.06	84.19	9.95
de	23518	2.48	89.24	60.04	82.6	10.73
jp	18637	1.97	86.74	56.22	84.25	9.89
uk	14338	1.51	94.09	72.98	91.92	14.31
br	14032	1.48	93.49	75.19	91.31	15.28
in	13620	1.44	82.96	57.37	75.48	11.01
ir	13564	1.43	82.7	53.99	64.26	6.25

Identification of Companies Receiving Data

The broader purpose of this study is to examine the extent to which the OBA industry is meeting the normative goals embedded in FIPPs by focusing on adherence to the crippled principles found in self-regulatory guidelines. Even though this ap-

proach predisposes the analysis to favor industry viewpoints, it merely follows the normative basis for the guidelines, rather than the deceptive and convoluted definitions of "notice" and "choice" which they promote. Thus, the first step is to figure out who the prominent companies collecting data are, the type of business model they represent, and determine if there is a likelihood that users may have knowingly interacted with them previously and thus be aware of their data collection practices.

The most striking finding of the million site audit is that 82.35% of websites initiate third-party HTTP requests to a Google-owned domain. While the competitiveness of Google is well known in search, mobile phones, and display advertising, its reach in the web tracking arena is unparalleled. The next company, Facebook, is found on a still remarkable 33.71% of sites, followed by Amazon on 17.15% of sites, Twitter with 13.96%, AppNexus with 12.92%, Aol with 12.17%, and Oracle with 11.00%. The reach of the top 25 companies is detailed in Table 4.2 and demonstrates that the final company, IPONWEB, collects user data on over 5% of all sites, itself no small feat given the size of the Alexa population.

The vast majority of the prominent data collectors are online advertisers. As Table 4.2 details, 20 of the top 25 companies are either involved in advertising or retail sales. Given that advertising and shopping happens online, all of these companies are practicing the techniques of OBA, and their data collection practices raise the entire host of privacy concerns detailed in the introductory chapter.

Three companies, Oracle, Adobe, and Acxiom, all provide some sort of business-

focused services such as consumer data which may be used for marketing or managing websites. Adobe is somewhat of an outlier in that their primary product which collects data from the web, "Marketing Cloud", is mainly focused on providing data to site owners and may therefore have a somewhat lower negative impact on user privacy.

The other prominent data collectors are all involved in hosting website content for other parties, and data they do collect is most likely used for enhancing such services rather than profiling users. Amazon Web Services (AWS) is the most prominent content host, and while Amazon does practice OBA using the amazon-adsystem.com domain, it appears that AWS data is isolated from such efforts.[5] Finally, it is worth noting that Cloudflare is typically viewed as a content host, but is now offering advertising focused services as well.

Company	**% Pages**	**Revenue**	**Consumer Services?**
Google	82.35	Advertising	Yes
Facebook	33.71	Advertising	Yes
Amazon	17.15	Hosting, Retail	Yes
Twitter	13.96	Advertising	Yes
AppNexus	12.92	Advertising	No
Aol	12.17	Advertising	Yes
Oracle	11.00	Business Services	No
Adobe	10.62	Business Services	Yes

[5]Overall, OBA accounts for roughly 3% of Amazon-directed traffic on the Alexa top 100,000 sites.

Company	**% Pages**	**Revenue**	**Consumer Services?**
The Trade Desk	10.35	Advertising	No
Yahoo	10.34	Advertising	Yes
StackPath	9.63	Hosting	No
Acxiom	9.03	Business Services, Data Broker	No
Cloudflare	7.46	Hosting, Advertising	No
MediaMath	7.23	Advertising	No
Yandex	7.07	Advertising	No
Automattic	6.89	Hosting	Yes
comScore	6.88	Advertising	No
Rubicon Project	6.86	Advertising	No
Spot X Change	6.83	Advertising	No
Lotame	6.62	Advertising	No
Neustar	6.37	Advertising	No
OpenX	6.1	Advertising	No
Turn	6.69	Advertising	No
PubMatic	5.58	Advertising	No
IPONWEB	5.33	Advertising	No

Table 4.2: Identification of Companies

Returning to the issues of "notice" and "choice", it is reasonable to assert that if a company provides consumer services, there is a non-zero probability a user may accede to the terms of service of a particular company during the process of using such services and that these terms may reveal third-party tracking. For example, a

user whose HTTP request data is captured by a Twitter "Share Button" may have a user account with the social networking site and be familiar with the policy of "Widget Data" which states that "We may tailor the Services for you based on your visits to third-party websites that integrate Twitter buttons or widgets". However, of the top 25 companies, only seven, Google, Facebook, Amazon, Twitter, Aol, Adobe, Yahoo, and Automattic (parent and host of Wordpress), have consumer-facing services. Thus, for the majority of the top data collectors, users are very unlikely to have heard of the company or have previously acceded to terms of service. This undercuts the likelihood they have been notified their data may be collected in the way Twitter notifies users. However, it is entirely possible a user may learn about data collection in other ways.

Disclosure of Companies in Privacy Policies

The industry solution for notifying users of data collection, the "AdChocies" icon, has been proven to be ineffective and confusing by prior research. However, a much more straight-forward means for users to learn about data transfers exist: reading the privacy policy of the site she or he is visiting.

As noted above, `webXray` is used by this study to identify the data flows occurring on websites in the Alexa top one million population. This analysis provides a record of the companies which receive data on the home page of each site. When `webXray` analyzes a page it attempts to extract a link to the privacy policy, and

if such a link is found, it may be used by `policyXray` to locate and extract the text of the policy for a given website. Using the tracking attribution analysis from `webXray`, `policyXray` is able to search the text of the policy in order to determine if the company or its subsidiaries is in any way mentioned in the policy. Further details of the process may be found in Chapter Three.

For this study, 184,666 unique privacy policies were successfully extracted from 204,412 sites, representing over 20% of all sites investigated. The reason there are fewer unique policies than sites is that many sites may share a common policy, as is frequent when a single publisher or corporate parent controls a collection of sites.

There are three reasons why a policy for a given site was not found. First, downloading and parsing the privacy policy can fail, this happened in 11% of cases (ie. there were 230,096 attempts made which captured 204,412 site's policies). Second, a page may not have a link to a privacy policy, itself a de-facto failure of "notice". Third, `policyXray` is currently limited to crawling English-language links to "privacy" policies and sites not using English will be missed. There is no way to differentiate between the second and third conditions; however, given the Alexa list contains sites from non-English regions, it is possible the language barrier is a primary cause for missing policies.

The overall impact of the limitations mentioned above is that the findings of this section of the study should not be viewed as a statistically representative sample of all privacy policies. Nonetheless, this corpus of policies is the largest collected

to date, and this is the first study to ever audit policies for disclosure of detecting tracking mechanisms and provides a sound basis for drawing conclusions about the nature of privacy policies on the web.

To evaluate if industry promises of "notice" are being practiced as promised, a common-sense approach of looking at the privacy policies has been taken. Using this approach, it is amply clear that notice is simply not occurring in the vast majority of cases. Across all of the policies and data flows attributed to 183 companies examined, users reading policies will only be notified of data flows in 10.70% of cases. More troubling, 53 companies are never mentioned in any policies, leaving it unlikely users will ever discover their data has been transferred to a third-party. While lack of disclosure is a wide-spread problem, a closer look at the top 25 companies in Table 4.3 reveals it is not uniform.

Company	**Disclosure (%)**
All companies	10.70
Google	33.38
Facebook	14.59
Amazon	3.51
Twitter	8.37
AppNexus	0.27
Aol	2.5
Oracle	3.83
Adobe	4.19
The Trade Desk	0.07

Company	**Disclosure (%)**
Yahoo	2.73
StackPath	0.02
Acxiom	0.08
Cloudflare	0.31
MediaMath	0.05
Yandex	0.89
Automattic	2.49
comScore	0.7
Rubicon Project	0.06
Spot X Change	0.03
Lotame	0.1
Neustar	0.03
OpenX	0.16
Turn	N/A
PubMatic	0.09
IPONWEB	0.03

Table 4.3: First Party Disclosure

Google is the company users are most likely to learn about during an evaluation of a website's privacy policy, and is disclosed in 33.38% of cases examined. Facebook, Twitter, Aol, Yahoo, Oracle, Adobe, and Automattic are all mentioned in at least one percent of cases. It is particularly damning to industry claims of effective "notice" that with the exception of Oracle, the most disclosed companies all have

consumer services and users may already be aware of data collection. The companies they are *least likely to know about* - those which have no consumer services - are also least likely to be mentioned in privacy policies. Even if such companies are disclosed, users may have no idea what the companies do with user data. From a reasonable standpoint, using a common-sense approach, it is clear that it is impossible for users to be given sufficient notice of data collection by the majority of companies.

In light of the extraordinarily low rates of disclosure, it is *highly unlikely users will ever gain access to the privacy policies which govern the use of their data*. However, in the rare event they do learn of a given company, they will find the policies incredibly time consuming and difficult to read.

Length and Readability of Policies

To determine the degree to which meaningful "notice" may be achieved by reading a given privacy policy for one of the top 25 companies receiving user data, the privacy policies of the companies were evaluated along length and readability metrics. To choose the most relevant policy for each company, the home page of the most prominent service was visited and any link to a privacy policy was followed (see the footnotes in Table 4.3 for details). The primary policy, and only the primary policy, was extracted. In some cases companies may have several policies, but it is unreasonable to expect users to wade though vast volumes of secondary and tertiary materials, so such policies were not counted. This is for a common-sense

reason: prior research has estimated that the "national opportunity cost for just the time to read policies is on the order of $781 billion" (McDonald and Cranor, 2008). Reading several policies for each company could conceivably put the sunken cost of time spent reading policies into the trillions of dollars.

Once the policies were chosen, three measures were recorded: the number of words in the policy, the Flesch-Kincaid Grade Level, and the Flesch Reading Ease score. McDonald and Cranor have previously studied the time impact of reading privacy polices and "assumed an average reading rate of 250 words per minute"; this study follows their lead in determining the amount of time needed to read a policy and updates their findings with a focus on the companies receiving the most data. Flesch-Kincaid Grade Level is a score which corresponds to the appropriate level of education (based on a U.S. K-12 scale) needed to understand the text. While not anchored to a grade-level, the Flesch Reading Ease scale ranges from 0-100 with 0 being most confusing and 100 being most clear. All of the company policies were in the English language, which these readability metrics are geared towards.

Of the 25 company policies analyzed, the longest, for Neustar, is 5,952 words long, would take nearly 24 minutes to read, and is written at a 14.5 grade-level, which is equivalent to three years of college. The shortest policy, with 282 words, is an outlier for it belongs to Amazon Web Services and essentially delegates privacy issues to the entities responsible for the sites rather than Amazon in its role as a

content host, which is fairly reasonable.[6] Of the companies whose primary revenue stream is derived from advertising, Yahoo has the shortest policy with 1,480 words, which would take nearly six minutes to read, and require one year of college to understand.

Company	**Words**	**Minutes to Read**	**Reading Ease**	**FK Grade**
Google	2748	10.99	38.6	13.6
Facebook[7]	1771	7.08	39.97	13.3
Amazon[8]	282	1.12	50.53	9.3
Twitter	3676	14.70	36.73	14.6
AppNexus[9]	3718	14.87	50.77	11.2
Aol	2722	10.88	36.42	14.7
Oracle[10]	3666	14.66	36.22	12.8
Adobe[11]	1665	6.66	29.38	15.3
The Trade Desk	5605	26.02	39.47	13.5

[6]StackPath is likewise a content host with a short policy of 752 words.

[7]For Facebook used the "complete policy".

[8]For Amazon, the AWS policy was used which delegates policy issues to clients: "AWS will not disclose, move, access or use customer content except as provided in the customer's agreement with AWS."

[9]For AppNexus the "Platform Policy" was used as opposed to the "Website Policy".

[10]For Oracle, the AddThis policy was used, per the site: "AddThis has been acquired by Oracle and will soon transition to the Oracle Privacy Policy. Click here to view the Oracle Privacy Policy and read your options for opt out and compliance issues under that policy. Please note that the provisions of AddThis's Privacy Policy below will remain active until the policy transition is complete."

[11]For Adobe the Marketing Cloud policy was used.

Company	Words	Minutes to Read	Reading Ease	FK Grade
Yahoo	1480	5.92	40.28	13.2
StackPath	752	3.00	41.19	12.9
Acxiom[12]	1880	7.52	29.79	15.2
Cloudflare [13]	1532	6.12	42.82	12.2
MediaMath	4816	19.26	38.55	13.9
Yandex[14]	1681	6.72	37.37	16.4
Automattic	1422	5.68	32.73	14.0
comScore[15]	1820	7.28	41.8	12.6
Rubicon Project[16]	1967	7.86	27.96	15.9
Spot X Change	2178	10.87	36.73	14.6
Lotame[17][18]	3131	12.52	29.59	15.2
Neustar	5952	23.80	31.41	14.5

[12]For Acxiom the Live Ramp policy was used.

[13]Cloudflare now has advertisement-oriented services: "Cloudflare's Firebolt benefits publishers, advertisers, and end users by improving an ad network's ability to serve ads faster, over a secure connection."

[14]For Yandex, the English policy was used.

[15]For comScore used Scorecard policy.

[16]For Rubicon used full policy for "ADVERTISING TECHNOLOGY PRIVACY POLICY"

[17]Lotame calls itself a DMP (Data Management Platform), which can power both advertiser and publisher needs, but ultimately this is an ad revenue service, so I am classifying under AdTech.

[18]Lotame does not accept DNT from IE: "If Lotame receives a "Do Not Track" signal from any browser other than Internet Explorer, Lotame will implement an opt-out"

Company	Words	Minutes to Read	Reading Ease	FK Grade
OpenX[19]	3380	13.52	27.96	15.9
Turn	4438	17.75	32.53	14.1
PubMatic	4369	17.47	19.94	18.9
IPONWEB	951	3.80	40.58	13.1

Table 4.4: Length and Readability of Policies

In terms of overall reading difficulty, the easiest advertiser policy was for AppNexus, which has a reading ease score in the "Fairly Difficult" range at 50.77 and a grade level score corresponding to a high school junior. The most difficult policy was for Yandex, which corresponds to two years of graduate school and has a reading ease score in the "Difficult" range. According to research cited by the U.S. Library of Congress, 50% of U.S. adults cannot read at an eighth grade level,[20] meaning that half of the population cannot understand any of the privacy policies to which they are frequently subject.

It can be said without further qualification that this analysis of policies reveals a gross failure of effective notice both in terms of the burden on user time and the fact that at least half of the population is unable to understand the policies. Given the

[19]OpenX DNT policy is oddly specific, "To opt out of OpenX's use of local storage or the browser cache to provide its services, please (1) use any tools provided by your browser to clear local storage and the browser cache, and (2) turn on any "Do Not Track" header setting offered by your browser"

[20]http://blogs.loc.gov/national-book-festival/

difficulties in accessing the policies in the first place, the "notice" aspect of "notice and choice" is a clear failure. The "choice" aspect fairs no better.

Respect for User Choices

As noted earlier, following cues from the FTC, the OBA industry has largely reduced the full spectrum of FIPPs norms to two: "notice" and "choice". As the above findings have made amply clear, the prospects for a user being properly notified of data transfer are minimal. However, it is possible that users could signal their "choice" not to be tracked in such a way that the user would not need to be aware of the data collection in the first place. This can be done by including data in the body of each HTTP request which tells entities receiving the request what the user has chosen. While industry offers a myriad of "opt-out" mechanisms, they are inconsistent, often rely on cookies and therefore are not permanent, and largely require the user to be notified that such options exist in the first place.

While industry failed to develop a universal opt-out mechanism, privacy researchers have done so, and the mechanism, "Do Not Track" (DNT), has been implemented in all major web browsers. According to the technical document, DNT provides a "means of allowing users to express their preferences about tracking, including to opt out of tracking some or all of the time" (Mayer and Narayanan, 2011). A user who does not consent to having her or his web browsing history stored by third parties (known as "tracking") may simply turn on the DNT setting.

DNT is a universal setting, it is consistent across browsers, it is reliable in that it requires no cookies, and it does not require the user to be aware of, or notified of, third-party data transfers. The only thing DNT does not do is provide a technical means to force data recipients to respect the user's choice; that is voluntary.

In essence, DNT appears to be the perfect solution for the industry: *companies do not even have to worry about notice*, explain unique "opt-out" mechanisms, or maintain the significant internal programming infrastructure need to manage bespoke "opt-out" cookies. For users, it is easy to set the option and forget about it. Indeed, the FTC has been highly supportive of the standard for these very reasons.

Prior to delving into the specifics of each data collector's stances on DNT, it is instructive to see how widespread mention of the standard is in the policies of the population of websites. While the standard is relatively new, it is already mentioned in 6.27% of site policies examined. Given the complexity of parsing natural language, it was not possible in this study to determine how many of these mentions are commitments to respect DNT rather than acknowledgements it will be ignored. Future work may address this challenge. Nonetheless, this is a fairly encouraging signal which deserves to be monitored in the future. Less encouraging is the industry response.

If industry is in fact committed to user "choice", and is acting in good faith, then all of the companies receiving user data would clearly mention DNT and their

respect for user choice in their privacy policies. However, only 7 of 25 company policies mention DNT at all. Of those 7, only Twitter unambiguously commits to respecting the setting, stating that "We honor the Do Not Track browser option to give you control over how your website visits are used to personalize your Twitter experience and ads". It is worth noting that even this statement focuses on data *use* rather than data *collection*. Oddly, Lotame offers qualified support for DNT: "If Lotame receives a 'Do Not Track' signal from any browser other than Internet Explorer, Lotame will implement an opt-out". The reason for Lotame's approach is likely that Internet Explorer at one time had DNT turned on by default, effectively making tracking an "opt-in" rather than "opt-out" signal. Aol's statement on DNT, "Aol currently does not take action in response to these signals", is representative of the 5 companies who mention, but do not respect, DNT.

Company	**% Pages**	**DNT Mention**	**DNT Honor**
All Site Policies	-	6.27%	-
Google	82.35	N	N
Facebook	33.71	N	N
Amazon	17.15	N	N
Twitter	13.96	Y	Y
AppNexus	12.92	N	N
Aol	12.17	Y	N
Oracle	11.00	Y	N
Adobe	10.62	N	N
The Trade Desk	10.35	Y	N

Company	% Pages	DNT Mention	DNT Honor
Yahoo	10.34	N	N
StackPath	9.63	N	N
Acxiom	9.03	N	N
Cloudflare	7.46	N	N
MediaMath	7.23	N	N
Yandex	7.07	N	N
Automattic	6.89	N	N
comScore	6.88	Y	N
Rubicon Project	6.86	N	N
Spot X Change	6.83	N	N
Lotame	6.62	Y	Y*
Neustar	6.37	N	N
OpenX	6.1	N	N
Turn	6.69	N	N
PubMatic	5.58	Y	N
IPONWEB	5.33	N	N

Table 4.5: Respect for Do Not Track

Some commentators have claimed that DNT has failed to protect user privacy.[21] Such talk is premature as any perceived failure is not in the standard, which is technically sound and supported in all major browsers. Rather, the failure is the

[21]One of many examples may be found here: http://www.zdnet.com/article/why-do-not-track-is-worse-than-a-miserable-failure/

behavior of industry entities who are refusing to respect, or even mention, the standard. The example of Twitter, and to a lesser degree Lotame, shows that it is not an undue burden to respect the user choices embodied in the standard. Aol and others claim that "There is no standard that governs what, if anything, websites should do when they receive [DNT] signals", so they refuse to follow it. This is clearly disingenuous as Aol's policy links to a Digital Advertising Alliance page which requires third-party cookie acceptance to facilitate an opt-out, which is simply a vastly less effective and more cumbersome means of signaling choice.

It is clear that there is no real commitment to "choice" by the OBA industry at large. The fact that so few companies mention the DNT standard in the first place suggests they feel the mere mention of it threatens the status quo which has proven immensely lucrative. The history of FIPPs suggests that the "choice" paradigm itself represents an erosion of the normative foundations of widely embraced data protection guidelines. The failure of industry to respect DNT is therefore a stunning example of putting profit before principle. However, mere negligence also erodes user privacy in the OBA space.

Security Practices

The security of personal data is a core component of nearly all interpretations of FIPPs. The original HEW principles state that "Any organization creating, maintaining, using, or disseminating records of identifiable personal data must assure

the reliability of the data for their intended use and must take reasonable precautions to prevent misuse of the data". Likewise, the 2009 FTC OBA guidelines assert that "Any company that collects and/or stores consumer data for behavioral advertising should provide reasonable security for that data". Finally, the industry-sponsored "Self-Regulatory Principles for Online Behavioral Advertising" contains the following passage: "Entities should maintain appropriate physical, electronic, and administrative safeguards to protect the data collected and used for Online Behavioral Advertising purposes". While there may be agreement on little else, there is a general commitment to data security in all these versions of FIPPs.

In the context of OBA there are two main technical factors involved in protecting user data: transport encryption and storage encryption. Simply put, encryption is a means of securing data so that only the entity possessing the "key" may read it; those who do not possess the key will only be able to access random data which reveals nothing.

Transport encryption applies to the way data is encoded when it moves from one computer to another over a network in order to prevent eavesdroppers from reading the data, be they coffee-shop hackers or government spies. For example, when a user visits the New York Times website and downloads an advertisement from a Google server, the user's connection to Google may be encrypted so that other parties may not see the contents of the connection. Likewise, *storage encryption* refers to how the data is stored once it gets to the Google server and protects the data if an

unauthorized entity were to gain physical or remote access to Google's databases at any time after the data has been received or subsequently processed. With today's computing power it is simple to implement transport and storage encryption; not doing so constitutes negligence.

From the outside it is impossible to verify if entities collecting user data through OBA are employing sufficient storage encryption without an independent auditing body. While guidelines exist, there are no central trusted regulators who audit the data storage practices of the OBA industry. However, it is possible to determine if transport encryption is being used by examining the network traffic generated when loading a page in order to determine if connections are made utilizing standard Secure Sockets Layer (SSL) connections.

When `webXray` is used to scan the sites for this study, the non-secure "HTTP" address of all pages is requested. In 19.13% of cases, `webXray` is redirected to a secure version of the home page. While pages transmitted over SSL should *only* have elements which are requested with SSL (failure to do so results in a warning), pages without SSL may have elements which may or may not have SSL. In order to be compliant with both cases, 51.54% of third-party elements use SSL. However, this relatively high number is undercut by the fact that only 1.15% of third-party cookies use the secure flag which requires SSL, meaning that third-parties generally facilitate the least amount of security possible rather than enforcing higher standards. Thus, the security of third-party elements is independent from the security of the main

site, and largely in the control of the advertising networks who provide the code.[22]

Company	**% SSL**
All Site Home Pages	19.13
All 1P Requests	18.43
All 3P Requests	51.54
Google	71.25
Facebook	88.68
Amazon	60.68
Twitter	85.34
AppNexus	30.77
Aol	32.44
Oracle	14.03
Adobe	56.66
The Trade Desk	35.54
Yahoo	27.97
StackPath	61.17
Acxiom	34.31
Cloudflare	55.90
MediaMath	19.62
Yandex	87.19
Automattic	65.68
comScore	40.65
Rubicon Project	31.50

[22]Advertisers may give web masters the "option" of utilizing encrypted connections, but a true commitment to security means no such option is given.

Company	**% SSL**
Spot X Change	64.76
Lotame	20.05
Neustar	30.44
OpenX	27.97
Turn	13.26
PubMatic	20.05
IPONWEB	40.48
Top 25 average	44.65

Table 4.6: Transport Security

Table 4.6 details the percentage of requests which were secured for each of the top 25 companies receiving user data. Given that even a single unencrypted request can expose sensitive user data to unauthorized parties, if all companies were compliant with both broader FIPPs norms as well as self-regulatory guidelines, the rate of encryption would be 100%. However, the highest rate of encryption, by Facebook, is only 88.68%. The worst encryption rate, by Turn, makes Facebook look remarkably good by comparison with only 13.26% of requests secured. On average, the top 25 recipients of third-party HTTP requests only encrypt 44.65% of requests. Once again, we see that industry practice falls far short of even the modest demands of self-regulatory bodies. Although it is impossible to audit the storage encryption practices used by these companies, it is certainly possible they are as poor as their

transport practices.

Conclusion

Starting from the promise of Fair Information Practice Principles grounded in normative values of individual respect and ending in the compromised state of privacy in the age of Online Behavioral Advertising, the past several decades have produced a gradual and profound erosion of privacy protections for citizens. One contributing factor to this situation is that the true spirit and values embedded in FIPPs were successfully co-opted by industry with the assistance of hapless regulators more interested in facilitating trade than protecting individual rights. However, even an adherence to the weakened version of FIPPs advocated by industry would be better than what exists today.

Simply put, the OBA industry fails to live up to a reasonable interpretation of either the concepts of "notice" or "choice". The web is rife with companies which siphon user data, and it is nearly impossible to use the web without having one's activities tracked. The majority of companies receiving user data are not known to users, their existence is not disclosed in the policies of most websites, their policies are both time consuming and difficult to read, they ignore simple choice mechanisms, and their security practices fail to protect users from malicious parties.

The current state of self-regulation fails to meet FIPPs, and provides scant benefits to users. The major beneficiary of self-regulation is the OBA industry,

who have co-opted and ignored FIPPs in order to deliver billion dollar profits to shareholders. Essential norms regarding fairness in the processing of personal data have been sacrificed for profit.

Thus, privacy on the web is barely regulated in theory and virtually unregulated in practice. This results in fundamental harms to basic privacy rights which have been adopted around the world. Such rights, however, tend to operate in the abstract. Due to the fact that the web so throughly dominates modern life, there are many ways by which the direct harms of OBA can be tied to specific contexts. The following chapters will further delve into how the failure of privacy regulation on the web results in harms to larger social institutions and values. The final chapter in this volume will return to the issue of regulation in order to determine if "notice and choice" can work, and if not, what could.

Chapter 5

Privacy Implications of Health Information Seeking on the Web

Health information has been regarded as sensitive since the time of the ancient Greeks. In the 5th century B.C., physicians taking the Hippocratic Oath were required to swear that: *Whatever I see or hear in the lives of my patients...I will keep secret, as considering all such things to be private* (National Institutes of Health, History of Medicine Division, 2002). This oath is still in use today, and the importance of health privacy remains universally recognized. However, as health information seeking has moved online, the privacy of a doctor's office has been traded in for the silent intrusion of behavioral tracking. This tracking provides a valuable vantage point from which to observe how established cultural norms and technological innovations are at odds.

When new technological practices impact established norms researchers should determine the exact nature of the impacts, and if they are malevolent, they should provide guidance on how such practices may be rectified (Nissenbaum, 2004). The norms of privacy in the medical context are well established, rooted in specific forms of harm, and are largely premised on the assumption that medical professionals are the source of medical information. However, the Pew Research Center has determined that 72% of adult Internet users in the U.S. now go online to learn about medical conditions in addition to speaking with trusted medical professionals (Fox and Duggan, 2013). Thus, from a normative perspective, this chapter endeavors to determine how traditional values of medical privacy fare when applied to the new reality of online information seeking.

The first section of this chapter, *Health Information Norms*, reviews established norms in the health context. These norms are designed to defend against two forms of harm: public disclosure of personal information and the tertiary use of medical data for discrimination against the ill. Individuals concerned about these harms may withhold medical information, resulting in inadequate treatment and the spread of disease. To combat these risks, professional guidelines and laws have been codified to protect doctor-patient confidentiality.

To understand how the shift to online information seeking has impacted health norms, the second section, *Research Findings and Impact*, uses the `webXray` software platform to evaluate a population of 80,142 pages related to health information

seeking. Pages are evaluated to determine the degree to which data flows expose private health information to private corporations, and how this may result in both public disclosure and tertiary commercial discrimination.

The third section, *Application of Extant Norms to New Practices*, takes a fresh look at established norms and legal guidelines in order to determine if new information practices are being sufficiently regulated by the extant strictures. In areas where there are gaps between norms and new practices, suggestions for resolution are made in the concluding section.

Overall, the chapter establishes that health norms are well defined, they are being violated to an astounding degree by online tracking, and extant protections fail to protect patient confidentiality online.

Health Information Norms

Across all areas of record-keeping, norms and laws regarding health information are among the oldest, most firmly defined, and well-enforced. The reason for this is that there are clear harms which come from improper health information disclosure, professional organizations have long-standing prohibitions on the abuse of patient data, and legal frameworks governing data use are well-defined.

According to Allen, medical confidentiality is valued as it prevents the harms associated with "shame and violations of modesty" and "frees individuals from the burdens of stigma, inequality and discrimination" (Allen, 2008). Indeed, the

harms of the inappropriate disclosure of medical information generally fall under the rubrics of public disclosure of personal information and the tertiary use of medical information for discriminatory purposes. The ultimate impact of these two harms is that the possibility of negative privacy outcomes will disuade individuals from revealing pertinent information to caregivers, thus harming individual and community health.

Medical information does not exist in a vacuum and it is often indicative of one's broader activities and habits. Thus, medical records may in fact be viewed as revealing numerous aspects of one's life, many of which would be embarrassing or harmful to reputation if disclosed.

Rindflesich observes that medical records may contain information on "fertility and abortions, emotional problems and psychiatric care, sexual behaviors, sexually transmitted diseases, HIV status, substance abuse, physical abuse problems, genetic predispositions to diseases, and so on" (Rindfleisch, 1997). Considering that stress is a factor in many medical conditions, it logically follows that the topics which cause the most mental anguish are also those which may be implicated in medical diagnosis and treatment.

Given both the intrinsic sensitivity of health conditions as well as the constellation of associated factors listed above, the release of medical information "can harm us...by causing social embarrassment or prejudice" (Rindfleisch, 1997). Furthermore, "individuals may experience...humiliation, shame, anxiety, and depres-

sion if their health secrets are revealed" (Rothstein and Talbott, 2006). Tragically, disclosure of one's anxiety and depression would likely greatly exacerbate such conditions. While shame and embarrassment may be viewed as harms which occur in the recesses of one's mind, the disclosure and tertiary use of health information may have external outcomes as well.

The second primary harm associated with the disclosure of medical information is that it may be repurposed for tertiary uses in order to discriminate against the ill who may be viewed as less reliable employees, financial risks, or otherwise undesirable.

While patients have an expectation that medical information "will be used only in the context of providing effective care", data may end up being used for "other purposes not envisioned in patient consent forms" such as data mining (Rindfleisch, 1997). Such tertiary use can "result in the inability to obtain insurance or employment" (Rothstein and Talbott, 2006). An additional factor is that due to the "commingling of the insurance and employment functions in the United States" some employers may have access to their employees' heath records and may seriously abuse "confidential medical information" (Starr, 1999). This risk is reflected in a survey which found that "52 percent of respondents said they were 'very concerned' or 'somewhat concerned' that insurance claims information might be used by an employer to limit their job opportunities" (Bishop et al., 2005).

Beyond discrimination in insurance or the workplace, commercial discrimination

may occur as a result of personal heath information being used for "marketing and other purposes" (Starr, 1999). Such discrimination is often overlooked as "the harm to the individual is ambiguous or relatively small" (Starr, 1999), yet these uses "conflict deeply with the confidentiality understandings most patients have when they sign consent forms" (Rindfleisch, 1997). Finally, due to the ubiquity of health record-keeping systems, "individuals may not be able to withhold sensitive health information from other unknown third parties if they want to be considered for employment, other essential life activities, or insurance" (Rothstein and Talbott, 2006).

According to Rindfleisch, there exists a paradox in the fact that "our medical records contain information about us that is of the utmost sensitivity, yet this information is only useful to us when it is shared with the medical providers and system under which we get our care" (Rindfleisch, 1997). Indeed, *withholding* information from caregivers may result in the unchecked spread of disease, increased costs of treatment, and even death. However, Allen observes that those "concerned about discrimination, shame or stigma have an interest in controlling the flow of information about their health" and may therefore have an interest in withholding information, even if it is ultimately harmful to their well-being (Allen, 2008).

Rothstein and Talbott note that "the mere possibility" that personal health information could be leaked "may lead individuals to forgo some potentially beneficial medical tests and procedures or even medical care altogether" (Rothstein

and Talbott, 2006). This problem is particularly troublesome among "vulnerable patients (e.g., minors) and individuals with potentially stigmatizing medical conditions (e.g., HIV infection, substance abuse, and mental illness)" (Rothstein and Talbott, 2006). Socially marginalized patients "may be even more likely to forgo care that is essential for their own health and the health of the public" (Rothstein and Talbott, 2006).

The risks are far from theoretical. One survey found that "one out of eight consumers has put their health at risk by engaging in such behaviors as: avoiding their regular doctor, asking their doctor to fudge a diagnosis, paying for a test because they didn't want to submit a claim, or avoiding a test altogether" (Bishop et al., 2005). At-risk groups such as the "chronically ill, younger, and racial and ethnic" minorities "are more likely than average to practice one or more" methods of withholding important medical information from caregivers (Bishop et al., 2005). Attempts at obfuscating medical information for fear of disclosure was "most common among those diagnosed with a disease", 15% of whom had withheld information (Bishop et al., 2005). Indeed, those "with a diagnosed disease are twice as likely as healthy consumers to ask a physician to obfuscate their diagnosis" (Bishop et al., 2005).

As the above has made clear, the twin specters of public disclosure and tertiary discrimination lead many to withhold information which may result in less effective medical treatment. Such withholding may harm both the patient and the wider

community. For this reason, strong professional guidelines and legal restrictions have been put in place to protect medical confidentiality.

The "foundational principle of medical ethics" is doctor-patient confidentiality (Rothstein and Talbott, 2006). This principle is represented in several long-standing professional norms. As noted above, the Hippocratic Oath has been in use since the 5th century B.C. and is still used today. Similar provisions were established in the "American Medical Association's first Code of Ethics in 1847" (Rothstein and Talbott, 2006). Likewise,the "American Psychological Association's ethical code requires that 'Psychologists respect . . . the rights of individuals to privacy, confidentiality, and self-determination'" (Allen, 2008). Similar confidentiality oaths exist for "nurses, dentists, and other health professionals" (Rothstein and Talbott, 2006).

While professional codes of ethics both indoctrinate and guide medical caregivers, they alone are not sufficient to protect patient privacy. For this reason, numerous legal statutes provide both enunciation of norms and define consequences for those who contravene them.

In the same way that the personal benefit of medical treatment is limited by the degree to which an individual discloses informative to a caregiver, the degree to which medical *institutions* may fulfill their purpose is limited by the trust put in them by patients. Thus, regulators "must not only protect privacy, health, and other legitimate interests" but they must also "produce public trust in institutions"

(Starr, 1999). Laws regarding health information are designed to protect not only individuals, but the very legitimacy of the medical establishment.

In the United States, "public concern about medical confidentiality is reflected in privacy rules promulgated under the Health Insurance Portability and Accountability Act (HIPAA)" (Allen, 2008). Pursuant to HIPAA, the U.S. Health and Human Services Department created the "Standards for Privacy of Individually Identifiable Health Information" which is also known as "The Privacy Rule" (Rothstein and Talbott, 2006). This rule governs how "3 classes of 'covered entities' (health providers, health plans, and health clearinghouses)" must protect patient health data (Rothstein and Talbott, 2006). Generally speaking, HIPAA provides numerous restrictions on how medical information may be shared, while still facilitating medical care, billing, and related tasks.

While HIPAA is fairly well-known, it is not the final word in medical privacy in the U.S.. The Americans with Disabilities Act of 1990, regulates "the flow of health information to employers so that individuals with disabilities may be considered for employment based on their abilities before an employer may consider an individual's actual or perceived limitations" (Rothstein and Talbott, 2006). Furthermore, federal medical privacy law does not "preempt more rigorous state patient privacy protections" of which there are many (Terry, 2003). For example, most "states have laws restricting the disclosure of HIV/AIDS information without individual consent" (Rothstein and Talbott, 2006).

Research Findings and Impact

The norms and restrictions described above trace their roots to antiquity, but the myriad changes brought about by the Internet represent essential challenges to the continued functioning of established guidelines. As detailed in the introductory chapter, when users visit a given website their browsing data may be transmitted to a range of unseen actors, many of which are collecting data in order to target advertisements to the user. In order to determine how these new data flows impact established norms we must select a population of health-related web pages, determine the scope of data flows, the recipients of data, the medical information being transmitted, and determine if the traditionally recognized harms of public identification and tertiary discrimination are to be found.

Population Selection

Establishing a population of websites which reflects the normative basis of the inquiry is essential for this research task. In this case, the objects of study are the websites' users searching for health information are most likely to visit. The Pew Internet and American Life Project found that "77% of online health seekers say they began at a search engine such as Google, Bing, or Yahoo" (Fox and Duggan, 2013) as opposed to a health portal like WebMD.com, thus a search-based approach to population construction is best.

In order to model the pages a user would visit after receiving a medical diagnosis,

a list of 1,986 diseases and conditions was compiled based on data from the Centers for Disease Control, the Mayo Clinic, and Wikipedia. Next, the Bing search API was used to find the top 50 search results for each term.[1] Once duplicates and binary files (pdf, doc, xls) were filtered out, a set of 80,142 unique web pages remained.

The pages collected reflect the diversity of sources for medical information including online support groups, newspaper articles, hospitals, and medical non-profits. A major contribution of this study to prior work is the fact that this analysis is focused on the pages which users seeking medical information are most likely to visit, irrespective of if the site is health-centric.

Data Collection

In April 2014, the 80,142 web pages described above were loaded with the `webXray` software in order to reveal top-level data on the flow of user data, the entities receiving data, and the degree to which specific health information was disclosed.

Top-Level Data Flows

Within the population studied, the flow of personal browsing data to third parties is pervasive across a number of measures. Based on information gleaned from the top level domains used by pages, five broad categories of pages have been investigated: all pages, commercial pages (.com), non-profit pages (.org), government

[1] Search results were localized to US/English.

pages (.gov), and education-related pages (.edu). Table 5.1 provides a break-down of findings across these five groups.

Of all pages examined, 91% leak data to a third party, 86% download and execute third party Javascript, and 71% utilize third party cookies. Unsurprisingly, commercial pages were above the global mean and had the most third party requests (93%), Javascript (91%), and cookies (82%). Education pages had the least third party HTTP requests (76%) and Javascript (73%), with a full quarter of pages free of data leakage. Government pages stood out for relatively low prevalence of third party cookies, with only 21% of pages storing user data in this way.

Table 5.1: Aggregate Trends

TLD	**N**	**% w/Request**	**w/JS**	**w/Cookie**
*	80,142	91	86	70
com	49,174	93	91	84
org	16,072	93	76	60
gov	7,991	88	86	19
edu	3,444	75	74	40

Recipients of Data Flows

While security and privacy research has often focused on *how* user privacy is violated, insufficient attention has been given to *who* is collecting user information. The simple answer is that a variety of advertising companies have developed a mas-

sive data collection infrastructure which is designed not only to avoid detection, but also ignore, counteract, or evade user attempts at limiting collection. Despite the wide range of entities collecting user data online, a handful of American advertising firms dominate the landscape.

78% of pages analyzed included elements which were owned by Google. Such elements represent a number of hosted services and use a variety of domain names: they range from traffic analytics (google-analytics.com), advertisements (doubleclick.net), hosted Javascript (googleapis.com), to videos (youtube.com). Regardless of type of services provided, in some way all of these HTTP requests funnel information back to Google. This means that a single company has the ability to record the web activity of a huge number of individuals seeking sensitive health-related information without their knowledge or consent.

While Google is the dominant actor, it is far from alone. Table 5.2 details the top 10 companies found as part of this analysis along with the rankings of two data brokers. In second place is comScore who are found on 38% of pages, followed by Facebook with 31%. It is striking that these two companies *combined* are still have less reach than Google.

Additionally, companies were categorized according to their type of revenue model. 80% of the top ten companies are advertisers who use web browsing data for behavioral targeting. The only exceptions to this rule are Adobe and Amazon. Adobe offers a mix of software and services, including traffic analytics. Amazon is in

Table 5.2: Corporate Ownership and Risk Assessment (N=80,142)

Rank	**% Pages**	**Company**	**Revenue**	**Identification**	**Discrimination**
1	78	Google	Advertising	X	X
2	38	comScore	Advertising	–	X
3	31	Facebook	Advertising	X	X
4	22	AppNexus	Advertising	–	X
5	18	Add This	Advertising	–	X
6	18	Twitter	Advertising	–	X
7	16	Quantcast	Advertising	–	X
8	16	Amazon	Retail & Hosting	–	X
9	11	Adobe	Software & Services	–	X
10	11	Yahoo!	Advertising	–	X
...	–	–	–	–	–
31	5	Experian	Data Broker	X	–
...	–	–	–	–	–
47	3	Acxiom	Data Broker	X	–

the business of both consumer-retail sales as well as web hosting with the Amazon Web Services (AWS) division. At present it is unclear if AWS data is integrated into Amazon product recommendations or deals, but the possibility exists.

While advertisers dominate online tracking, two major data brokers were also detected: Experian (5% of pages), and Acxiom (3% of pages). The main business model of data brokers is to collect information about individuals and households in order to sell it to financial institutions, employers, marketers, and other entities with such interest. "Credit scores" provided by Experian help determine if a given individual qualifies for a loan, and if so, at what interest rate. Given that a 2007 study revealed that "62.1% of all bankruptcies...were medical" (Himmelstein et al., 2009), it is possible that some data brokers not only know when a given person suffered a medical-related bankruptcy, but perhaps even when they first searched for information on the ailment which caused their financial troubles.

Health Information Leakage

The HTTP 1.1 protocol specification warns that "the source of a link [URI] might be private information or might reveal an otherwise private information source" and advises that "[c]lients SHOULD NOT include a Referer header field in a (non-secure) HTTP request if the referring page was transferred with a secure protocol" (Fielding et al., 1999). In simpler terms, web pages that include third party elements, but do not use secure HTTP requests, risk leaking sensitive data via the "Referer" field.

Of the pages analyzed, only 3.24% used secure HTTP, the rest used non-encrypted HTTP connections and thereby potentially transmitted sensitive information to third parties. Unsurprisingly, a significant amount of sensitive information was included in URI strings.

Based on a random sample of 500 URIs taken from the population of pages analyzed (N=80,142), 70% contained information related to a specific symptom, treatment, or disease. An example of an URI containing specific symptom information is:

`http://www.nhs.uk/conditions/breast-lump/[...]`

a URI containing no such information is:

`http://www.ncbi.nlm.nih.gov/pubmed/21722252`

Given that the former type of URI was by far the most prevalent, it may be seen that third-parties are being sent a large volume of sensitive URI strings which may be analyzed for the presence of specific diseases, symptoms, and treatments. This type of leakage is a clear risk for those who wish to keep this information out of the hands of third-parties who may use it for unknown ends.

Impact of Tracking on Established Harms

As noted above, the release of personal health information carries with it a risk of two interrelated harms. The first is *public disclosure of personal information*, where

an individual's name is publicly associated with their medical history. The second is *tertiary use and discrimination*, where an individual's name is not necessarily revealed to the public, but they may be discriminated against based on perceived medical conditions.

While many people may consider details of their health lives to be of little interest or value to others, such details form the basis of a lucrative industry. In 2013, the US Senate Committee on Commerce, Science and Transportation released a highly critical review of the current state of the so-called "data broker" industry. Data brokers collect, package, and sell information about specific individuals and households with virtually no oversight. This data includes demographic information (ages, names and addresses), financial records, social media activity, as well as information on those who may be suffering from "particular ailments, including Attention Deficit Hyperactivity Disorder, anxiety, depression...among others" (Staff of Chairman Rockefeller, 2013). One company, Medbase200, was reported as using "proprietary models" to generate and sell lists with classifications such as "rape sufferers", "domestic abuse victims", and "HIV/AIDS patients" (Dwoskin, 2013).

It should also be noted that such models are not always accurate. For example, individuals looking for information on the condition of a loved one may be falsely tagged as having the condition themselves. This expands the scope of risk beyond the patient to include family and friends. In other cases, an individual may be searching for health information out of general interest and end up on a data broker's

list of "sufferers" or "patients". Common clerical and software errors may also tag individuals with conditions they do not have. The high potential for such errors also highlights the need for privacy protections.

Furthermore, poorly protected health information may be publicly leaked by criminals. The retailer Target has used data-mining techniques to analyze customers' purchase history in order to predict which women may be pregnant in order to offer them special discounts on infant-related products (Duhigg, 2012). Even if shoppers and surfers are comfortable with companies collecting this data, that is no guarantee it is safe from thieves. In 2013, 40 million credit and debit card numbers were stolen from Target (Krebs, 2013). While a stolen credit card may be reissued, if Target's health-related data were leaked online, it could have a devastating impact on millions of people. Merely storing personally identifiable information on health conditions raises the potential for loss, theft, and abuse.

Advertisers regularly promise that their methods are wholly anonymous and therefore benign, yet identification is not always required for discriminatory behavior to occur. In 2013, Latanya Sweeney investigated the placement of online advertisements which implied a given name was associated with a criminal record (Sweeney, 2013). She found that the presence of such ads were not the result of particular names being those of criminals, but appeared based on the racial associations of the name, with "black names" more often resulting in an implication of criminal record. In this way extant societal injustices may be replicated through

advertising mechanisms online. Discrimination against the ill may also be replicated through the collection and use of browsing behavior.

Data mining techniques often rely on an eclectic approach to data analysis. In the same way that a stew is the result of many varied ingredients being mixed in the same pot, behavioral advertising is the result of many types of browsing behaviors being mixed together in order to detect trends. As with ingredients in a stew, no single piece of data has an overly deterministic impact on the outcome, but each has *some* impact. Adding a visit to a weather site in the data stew will have an outcome on the offers an user receives, but not in a particularly nefarious way. However, once health information is added to the mix, it becomes inevitable that it will have some impact on the outcome. As medical expenses leave many with less to spend on non-essential items, these users may be segregated into "data silos" of undesirables who are then excluded from favorable offers and prices (Turow, 2012). This forms a subtle, but real, form of discrimination against those perceived to be ill.

Having collected data on how much tracking is taking place, how it occurs, and who is doing it, it is possible to evaluate the impact this has on users. Table 5.2 shows a breakdown of how data collection by twelve companies (top ten and data brokers) impacts the risks of both public disclosure and tertiary discrimination. The two data brokers most obviously entail a personal identification risk as their entire business model is devoted to selling personal information. It is unlikely they

are selling raw web tracking data directly, but it may be used as part of aggregate measures which are sold.

Despite the fact that Google does not sell user data, they do possess enough "anonymous" data to identify many users by name. Google offers a number of services which collect detailed personal information such as a user's personal email (Gmail), work email (Apps for Business), and physical location (Google Maps). For those who use Google's social media offering, Google+, a real name is forcefully encouraged. By combining the many types of information held by Google services, it would be fairly trivial for the company to match real identities to "anonymous" medical-related web browsing data. Likewise, Facebook requires the use of real names for users, and as noted before, collects data on 31% of pages; therefore, Facebook's collection of browsing data may also result in personal identification. In contrast, Twitter allows for pseudonyms as well as opting-out of tracking occurring off-site.

The potential for tertiary discrimination is most pronounced among advertisers. As noted above, online advertisers use complex data models which combine many pieces of unrelated information to draw conclusions about "anonymous" individuals. Any advertiser who is collecting and processing health browsing data will use it in some way unless it is filtered and disposed of.

Application of Extant Norms to New Practices

Having established what the norms and laws are, and what the online practices are, it follows that we must determine if extant protections guard against new harms. Unfortunately, they often fall far short of what is needed.

As noted above, HIPAA is the best known health privacy law in the United States. However, HIPAA is not a panacea and is closely tailored to overseeing those providing health-specific services such as doctors, hospitals, and insurance claims processors. According to Terry, while "the HIPAA privacy regulations do protect patients in many of their interactions with bricks-and-mortar providers", "the same cannot be said for most consumer interactions with online providers" (Terry, 2003). Put another way, HIPAA stops at the door of the doctor's office.

Nevertheless, the U.S. Federal Trade Commission (FTC) has established a "Health Breach Notification Rule" which requires entities holding personally identifiable health records to notify users if such records have been stolen (US Federal Trade Commission, 2010a). However, merely providing health information (rather than storing doctor's notes or prescription records) does not place a business under the jurisdiction of HIPAA or associated rules. Many businesses that handle health information are subject to virtually no oversight and the main source of policy regarding the use of health information online comes in the form of self-regulation by the parties which stand to benefit the most from capturing user data: online advertisers.

In the same way that professional norms in the medical profession govern how doctors respect patient confidentiality, in theory industry self-regulations could provide similar protection. However, self-regulation in the online advertising space has proven wholly insufficient to produce normatively acceptable medical privacy outcomes. The FTC determined that "industry efforts to address privacy through self-regulation have been too slow, and up to now have failed to provide adequate and meaningful protection" (US Federal Trade Commission, 2010b).

A close reading of industry guidelines shows that there are no serious sanctions for violating the rules which advertisers draw up among themselves. The Network Advertising Initiative (NAI) has produced a "Code of Conduct" which requires opt-in consent for advertisers to use "precise information" about health conditions such as cancer and mental-health (Network Advertising Initiative, 2011). Yet the same policy also states that "member companies may seek to target users on the basis of such general health categories as headaches" (Network Advertising Initiative, 2011). Given that the range of ailments between cancer and a headache is incredibly broad, this directive provides virtually no oversight.

Likewise, the Digital Advertising Alliance (DAA) provides rules which also appear to protect health information, but legal scholars have determined that "an Internet user searching for information about or discussing a specific medical condition may still be tracked under the DAA's principles" (Hoofnagle et al., 2012).

Overall, while the harms remain the same, the normative frameworks which

have previously controlled medical information simply do not exist in the online context. Once again it is clear that self-regulation is not a replacement for more robust forms of regulation.

Conclusion

At this point it should be clear that the transition to web-based medical information seeking has had deleterious effects on medical privacy. This transition has introduced a range of harms to patient privacy without the protections offered by professional medical guidelines or most laws. The final chapter of the dissertation will present an overview of approaches to regulation which could bring the current practices of online medical information seeking in harmony ancient norms of medical privacy. First, it is necessary to interrogate how OBA has a negative impact on democratic norms.

Chapter 6

Privacy, Trust, and Security Implications of Behavioral Advertising on News Websites

[1]

The first case-study firmly established that the current state of self-regulation of online behavioral advertising (OBA) is failing to meet both broader social and legal norms related to automated information processing, as well as the minimum guidelines related to the "notice and choice" framework advocated by industry. The second case-study demonstrated the ways in which the lack of regulation in OBA has led to deep personal and societal harms in the realm of medical privacy. The

[1]Thanks to Bo Mai for inspiration for the chapter title.

third and final case study focuses on the ways in which OBA harms the body politic by undermining the ability of the press to fulfill its democratic purpose.

Prior research has noted that news websites contain vastly more behavioral tracking mechanisms than other types of sites (Englehardt and Narayanan, 2016; Budak et al., 2016). The main reason for this discrepancy is that news production in the U.S. has long been a predominately commercial enterprise which is funded by a mix of advertising and subscription revenue (Pickard, 2014). In the move to online publishing, traditional sources of subscription and advertising revenue have rapidly evaluated, leading to a search for new forms of revenue. One of the most prominent new forms of revenue for online publications has come from the dominant business model of the web: online behavioral advertising (OBA) (Mitchell and Rosenstiel, 2015). It is therefore fairly clear why news websites have more tracking, what is unclear is how the shift to OBA has impacted the relationship between citizens and the press.

Beyond entertainment and sports, the press provides citizens with vital information on the functioning of government. For this reason, the U.S. Constitution provides for broad press freedoms and the press is viewed as a core component of the democratic system of governance, often called the "fourth estate". There are three primary ways in which the press serves a democratic purpose which are pertinent to the topic of OBA. First, as an independent social institution the press functions best when it is free from undue influence and serves the interests of citizens above

all else. Second, the press is ideally devoted to upholding values of transparency and accuracy in order than citizens may place trust in news media. Third, and most important, the press provides the mechanism for facilitating public deliberation on matters of governance. The ultimate check on power is exercised by citizens in the voting booth, and the press is the means by which citizens are able to make democratic choices and restrain tyranny.

Online behavioral advertising compromises the above functions in several ways. First, while press outlets require independence to operate without influence, OBA fosters a centralization of both revenue and content-delivery infrastructure, which gives a handful of OBA firms massive unseen leverage over the press. Second, the press is unable to function without the trust of citizens. However, the essential nature of OBA is premised on extracting user data in covert ways which may expose users to criminal malware, neither of which inspires trust. Finally, the free and open airing of facts and ideas which the press fosters is threatened by OBA as OBA provides a means for the government to surveil citizens, which may result in a chilling effect whereby citizens are unable to freely participate in democratic functions.

To determine if OBA is in fact undermining the social-normative function of the press, a population of nearly 250,000 pages drawn from over 5,000 U.S.-based news websites is analyzed across several dimensions. First, in order to establish the extent of OBA in the sector, a broad census of tracking mechanisms and base-

line privacy impacts is conducted. Second, in order to explore the independence of the press, an analysis of underlying data flows is conducted to determine if publishers operate independently or are reliant on centralized revenue and hosting infrastructures. Third, to determine if OBA undermines the trustworthiness of news websites, privacy policies are evaluated to determine if they are written in an easily understandableway and if they fully disclose the companies which harvest user information from a given site. Additionally, web pages are checked to see if they further undermine user trust by partnering with advertising networks which have hosted malware. Finally, using public disclosures of government mass-surveillance as a starting point, the extent to which the government may surveil visitors to news websites and chill deliberation is explored.

The analyses conducted in this chapter demonstrate that the high concentration of OBA mechanisms on news websites is not a mere curiosity or artifact of commerce, it is a threat to the foundation of democracy. The press cannot put the needs of citizens first if they are beholden to a handful of OBA companies which provide nearly all revenue and content delivery infrastructure. The press cannot maintain the trust of citizens if it engages in unsavory and deceptive advertising practices. Finally, the press cannot act as a check on power if its embrace of OBA gives government agencies a means to surveil the reading habits of citizens.

This chapter will begin with a background of the pertinent normative concerns related to the press. Second, ways in which OBA may negatively impact these values

will be detailed. Third, a description of the construction of the page population used for the study as well as particulars related to methodological choices which are unique for this chapter is provided. Fourth, research findings across several dimensions are presented. Finally, the conclusion will provide brief reflections on how OBA negatively impacts the press.

Normative Background

It is widely recognized that a free press is an essential component of a democratic society. There are many factors which account for both the ways in which a free press contributes to a democratic society as well as the means by which it may be successful in these tasks. Three are particularly relevant to the effects of online behavioral advertising (OBA) on news websites. First, the press must be viewed as independent and representing the interests of citizens above all else. Second, citizens must view the press as being truthful and accurate in order to create a foundation of trust. Third, the press's primary democratic function is to facilitate public deliberation on matters of governance. However, this deliberative function may be negatively impacted by the "chilling effects" of surveillance.

It is widely recognized that the press must be independent of outside control in order to function as a check on power. According to Kovach and Rosenstiel the "first loyalty" of the press it to citizens, therefore journalists "must maintain an independence from those they cover" and "serve as an independent monitor of

power" (Kovach and Rosenstiel, 2007). Likewise, Deuze states that "editorial independence" is a core "ideological value" of the press (Deuze, 2005). This viewpoint is echoed by Bennett who states that the concept of press independence is "central to the organization, make-up, working practices and output of media systems across the globe" (Bennett et al., 2015). Karppinen and Moe summarize the dominant perspectives when they state that independence is a "central normative principle in media policy and journalism" (Karppinen and Moe, 2016)

The concept of press independence is often premised on freedom from external forms of control on reporting and publishing. Bennet observes that "media independence has come to mean working with freedom: from state control or interference, from monopoly, from market forces, as well as freedom to report, comment, create and document without fear of persecution" (Bennett et al., 2015). Deuze states that independence is a value held closely by "reporters across the globe [who] feel that their work can only thrive and flourish in a society that protects its media from censorship; in a company that saves its journalists from the marketers" (Deuze, 2005). Freedom from commercial influence is however put at risk by "native advertising and other practices online that blur the line between journalism and sponsored content" and threaten "the fundamentals of journalistic independence" (Karppinen and Moe, 2016). Overall, the true value of independence is it allows journalists to put the needs of citizens first and pursue the truth.

Beyond being seen as independent, it is self-evident that citizens must trust

news media outlets if journalism is to fulfill a positive democratic function. In *The Elements of Journalism*, Kovach and Rosenstiel state that journalism's "first obligation is to the truth" and that the "essence" of journalism "is a discipline of verification" (Kovach and Rosenstiel, 2007). This adherence to truth is often portrayed as "objectivity". Deuze notes that scholars "have identified objectivity as a key element of the professional self-perception of journalists" and an "ideological cornerstone of journalism" (Deuze, 2005). This notion has a long history, and Boudana notes that the concept of objectivity was "first used by journalists in the United States and Britain within the framework of the promotion of professionalism" and "surfaced as a norm in the 19th-century" (Boudana, 2011). Furthermore, the underlying democratic purpose of truthful reporting is clearly spelled out in a handbook from the Society of Professional Journalists which states that "accuracy and fairness speak to the obligation of providing meaningful information to citizens who depend on quality, authenticity, and lack of bias to understand issues and to make important decisions" (Black et al., 1999). Thus, accurate information may lead to productive discussions on governance by citizens.

The ultimate democratic purpose of the press "is to provide citizens with the information they need to be free and self-governing" and to "provide a forum for public criticism and compromise" (Kovach and Rosenstiel, 2007). This forum is often referred to as the "public sphere" and the process of "criticism and compromise" is often referred to as "deliberative democracy". Habermas states that this

"deliberative paradigm...is supposed to generate legitimacy through a procedure of opinion and will formation that grants (a) publicity and transparency for the deliberative process, (b) inclusion and equal opportunity for participation, and (c) a justified presumption for reasonable outcomes (mainly in view of the impact of arguments on rational changes in preference)" (Habermas, 2006) Christians puts this sentiment in simpler terms, stating that the news media have a "facilitative" role in which they "support and strengthen participation in civil society outside the state and market" in order to "promote dialogue" among citizens which leads to the formation of public opinion (Christians, 2009). However, journalists and citizens must feel free to engage in deliberation for this function to be effective.

The deliberative function of the press is under constant threat from what is known as chilling effects. Chilling effects refers to a broad theory that state surveillance "may chill or deter people from exercising their freedoms or engaging in legal activities" (Penney, 2016). Chilling effects can impact the deliberative function of the press both by restricting the activities of journalists as well as impacting the ability of the public to freely seek information and discuss controversial topics.

An extensive study by the PEN America Center conducted in response to revelations of mass spying by the N.S.A. (details of which are discussed later) found that 93% of journalists were very concerned over government efforts to unmask journalistic sources. The report concluded that government "surveillance will damage the ability of the press to report on the important issues of our time if journalists

refrain from contacting sources for fear that their sources will be found out and harmed, or if sources conclude that they cannot safely speak to journalists and thus stay silent" (America, 2013). Thus surveillance may limit the ability of the press to present controversial topics to the public.

There is also strong evidence of surveillance-linked chilling effects in the realm of information seeking and public deliberation. One study investigated if fear of N.S.A. surveillance would negatively impact citizens' willingness to seek out controversial materials online. Based on examination of search trends before and after revelations of N.S.A surveillance it was found that "there is a chilling effect on search behavior from government surveillance on the Internet" (Marthews and Tucker, 2015). Even if citizens do come across controversial material, experimental research has found that when Internet users are exposed to a fictional news story on U.S. military action and are primed to be cognizant of "government surveillance", the "likelihood of speaking out" on the topic was "significantly reduced" (Stoycheff, 2016). Thus, while the press fosters deliberation, both the production and consumption of journalism may be impacted by chilling effects linked to surveillance.

Research Questions

Based on the normative values associated with journalism detailed above, there are three attendant areas where the practices of online behavioral advertising (OBA) may have a negative effect on the efficacy of the press. First, by centralizing both the

basis for revenue generation and content hosting in a handful of companies, the press may lose significant autonomy and independence. Second, the opaque nature of OBA combined with its facilitation of criminal malware schemes may contravene the basis for trust in press outlets. Third, due to the fact that government intelligence agencies may leverage OBA tools to conduct mass surveillance, the role of the press as a facilitator of public deliberation may be undermined.

The internet has been characterized as a decentralized network which distributes media power away from legacy intermediaries and into the hands of the public writ large. However, the rise of a handful of giants in search (Google) and social media (Facebook, Twitter), shows that instead of removing intermediaries, the Internet has simply produced new ones whose power dwarfs that of former intermediaries. When it comes to the role of the press, a move to the web does not necessarily equate with increased independence, rather the dominance of OBA may in fact reduce the underlying independence publishers have enjoyed for centuries.

Publisher independence may be undermined if a small group of organizations controls the underlying revenue generation function of the press or if a small group controls the publishing infrastructure which is now composed of servers and data centers rather than printing presses. If such centralization exists, the press has little leverage to dictate policies related to the processing of user data, and they will end up beholden to powerful entities in a way which has the potential to impact the journalistic mission.

Two research questions may therefore by pursued: first, "How centralized, or distributed, are revenue generating mechanisms on news websites?", second, "How centralized, or distributed, is the hosting of content on news websites?".

More than ink and paper, the press has always relied on the trust of readers in order to thrive. Reader trust is first and foremost grounded in the degree to which news organizations provide full and fair accountings of relevant events. However, the technical underpinnings of OBA rely on covert surveillance of users' web browsing habits and as was seen in Chapter Four, full and fair accounting of web tracking practices is seldom found. Furthermore, the adoption of OBA puts users at risk of criminal attacks using so-called "malvertising" schemes, further undermining the basis for trust.

In order to determine if the shift to OBA has altered the normative basis for trust in news media, two additional research questions may be asked: first, "Are the privacy policies of news websites easy to understand and do they clearly and accurately disclose data flows to third-parties?", and two, "Should readers trust news websites to keep them safe from cyber attacks?"

As noted above, the normative and social value of the press is rooted in its ability to foster public deliberation on matters of governance. Likewise, the means by which the press acts on a check on power is that a "watchdog" press may expose government wrongdoing to citizens who may punish corrupt government authorities in the voting booth. However, OBA may undermine this role.

OBA techniques are by nature designed to centralize the collection of reader habits into corporate-controlled databases. Such corporations may be exploited or coerced into giving access to web browsing data to government intelligence agencies. Likewise, these companies are free to sell the information as they see fit, be it to marketers or the F.B.I. If news consumers feel they are being monitored they may be less likely to read controversial materials, visit news websites which offer an adversarial take on the actions of the government, or discuss controversial matters with other citizens. The final research question is therefore, "How may OBA chill deliberation and hinder the ability of the press to act as a check on, as opposed to a facilitator of, state power?".

Prior research has noted that news websites tend to have more tracking mechanisms that other websites (Englehardt and Narayanan, 2016; Budak et al., 2016). In order to add to existing knowledge on the topic, and to highlight the findings of this study, the frequency of tracking mechanisms on news websites is compared to a reference population in order that another research question may be addressed: "Are the privacy practices of news websites demonstrably worse than those of other websites?".

Methods and Population Design

This chapter uses both `webXray` and `policyXray` in order to collect and analyze data, details of which are found in Chapter Three. Considerations regarding the

design of the news website population, a reference population, the privacy policy corpus, as well as measures for malware exposure, are further explicated below.

The goal of this analysis is to provide an overview of the online news industry in the United States. To do so, creating a population of pages with two features was necessary: first, the population must have as large a number of distinct news websites as possible, and second, for each site, many pages must be analyzed to capture the greatest possible number of tracking elements. Both factors were successfully accomplished through a two-stage process of collection and expansion.

To build the initial population of sites, both the DMOZ and Alexa popular lists were consulted.[2] However, these lists proved both out-of-date and poorly organized. Instead, the website USNPL.com (U.S. News Paper List) was found to provide a well organized and up-to-date list of not only newspapers, but news-related magazines, television, and radio stations. Using the USNPL site resulted in a set of roughly 9,600 links to news-related websites.

To create a population of pages drawn from the above sites that were specific to news content, the Bing Search API was used to select 50 pages from each site containing the term "news". Not all of the sites identified from the USNPL produced 50 results as desired. Among a variety of factors accounting for this was that many of the pages hosted PDF versions of entire issues of papers rather than distinct pages for articles, some of the sites were no longer operational (potential victims of

[2]See Chapter Three for details on DMOZ and Alexa.

small-town newspapers closing), and some of the magazines listed were not news-oriented. Nevertheless, over 5,000 of the sites generated 50 pages with the term "news", generating a total population of 250,000 pages which were analyzed by `webXray`.

The computer used to collect data was based at the University of Pennsylvania, and data collection was performed between the dates of December 31, 2016 and January 15, 2017. The relatively long collection period is a reflection of a conscious choice to crawl the sites more slowly in order to reduce load on the analyzed sites and to not be flagged as a "bot".[3]

In addition to measuring the impacts of OBA on news websites, it was also necessary to establish a reference population so that it could be determined if news websites represented a greater risk to privacy than other popular websites. As discussed in Chapter Three, the top 100,000 Alexa sites are chosen as a reference population in order to compare news findings to what is normal for other popular websites. Again, using a computer based at the University of Pennsylvania, results were collected during the period of January 19, 2017 through January 20, 2017, providing a snapshot taken within days of the news sample, thereby establishing a sufficiently comparable set of pages.[4]

Once the `webXray` analysis was complete, `policyXray` was used to collect and

[3]Such an approach is generally viewed as being "polite" as well.

[4]The far more rapid rate of collection was due to each page belonging to a distinct site, thus there were no issues in regards to the "politeness" factor.

analyze privacy policies for the population of news pages and reference pages. As noted in Chapter Three, during page collection, `webXray` searches for the first link containing the English-language string "privacy policy", and falls back to the string "privacy". `policyXray` then takes these links and attempts to load, parse, and analyze the privacy policy.

For the population of U.S. news websites, 2,881 privacy policies corresponding to 140,340 pages were successfully extracted. These policies cover 3,125 of 5,095 distinct sites meaning that nearly 40% of news pages did not have a clear machine-readable English-language link to a privacy policy. For the reference population, 24,478 privacy policies corresponding to 26,758 sites were successfully extracted (28.48% of the total). The lower rate of success for the reference population may be partially attributed to the fact that the Alexa list contain non-English websites. Both corpuses are significantly larger than have ever been studied before and represent a significant scholarly contribution.

As detailed in Chapter Three, it is possible to extend `webXray` findings by querying outside data sources such as VirusTotal. In order to provide additional depth on the analysis of malware exposure, VirusTotal was queried for both the news and reference populations.

Findings

Across all dimensions examined, online behavioral advertising (OBA) on news websites has a negative impact on established social norms related to journalism. Furthermore, when compared to a reference population, news websites often have significantly worse effects than other popular websites.

Prevalence of OBA and Top-Level Privacy Impacts

Simply put, the OBA model enjoys wide adoption by online news websites and readers may expect to have their privacy compromised to an astounding degree. Fewer than five percent of news pages are encrypted, compared to nearly a quarter of non-news pages. Nearly all news pages examined transmitted user information to third-parties in the form of HTTP requests, nearly nine in ten contained a third-party cookie which may be used for tracking user behavior, and over nine in ten pages contained third-party Javascript code which may be used for advanced fingerprinting techniques.[5]

Table 6.1: Top-Level Privacy Findings

Population	N=	% Encrypted	% w/3P Request	% w/3P Cookie	% w/3P Javascript
News	244,293	4.45	97.88	89.33	96.06
Reference	93,926	24.28	92.62	77.83	89.15

[5]See Chapter Two for references to papers which discuss Javascript-based tracking.

In comparison to the reference population, the percentage of sites containing third-party requests, cookies, and Javascript on news websites is higher, but not overwhelming so as OBA is widespread across the web. However, these top-level metrics only indicate if data leakage is occurring *at all*, as opposed to how many requests are being made per-page (e.g. a page with 20 third-party requests and a page with one such request will be weighted the same in the above measures).

When accounting for the number of requests and cookies per-page, news sites are considerably worse than the reference population. News sites contact over twice as many third-party domains on average (41.43 vs. 17.92), and the mode of domains requested is a full order of magnitude greater (20 vs. 2). Likewise, the number of third-party cookies set is considerably greater: 27.33 on average for news compared to 11.14 for non-news. Thus, while both news and non-news sites practice OBA, visitors to news websites are exposed to vastly more tracking.

Table 6.2: Number of Distinct Third-Party Domains Requested and Cookies per Page

Population	Domains Requested Mean	Domains Requested Mode	Cookies Mean
News	41.43	20	27.33
Reference	17.92	2	11.14

It should be pointed out that the reference population only looks good in comparison to news websites. The findings across the reference sites indicate wide-

spread and deep privacy issues, not a "best practice" scenario. The fact that news sites under-perform the reference population to such a degree underscores just how exceptionally troublesome news websites are from a privacy perspective. In either case, the outlook for user privacy is poor, but from a perspective of independence of news media, the outlook is even more grim.

Centralization of Revenue and Content Delivery Infrastructure

Ownership consolidation in the news media industry is not a new phenomena. The days when local newspapers maintained private lists of subscribers, worked directly with local businesses to field advertisements, and owned printing presses are long gone, if they ever truly existed. Yet on a spectrum of possibility, such a distributed system represents one extreme where subscriber privacy and publisher independence are highest. One the other extreme, a small cartel could have records of all subscribers, broker all advertisements, and operate a handful of printing presses which would publish all papers. In the first extreme, independence would be highest, in the second newspapers would have so little independence that they would be demoted to content providers. By examining the centralization of data flows, revenue, and hosting, this study demonstrates that the truth is closer to the second extreme, and the spread of OBA is eroding the independence of the press in a way which has not been previously quantified.

Table 6.3 shows the top ten companies which receive user data on both news and reference web pages. The most remarkable finding is that while pages from over 5,000 distinct news outlets were studied, one company, Google, has the ability to monitor readers on nearly 95% of pages. Facebook, a company increasingly positioning itself as a news source, collects user data on 60.45% of pages. Beyond Google and Facebook, two other companies have the ability to monitor over half of pages: Oracle (55.06%) and Amazon (51.39%).[6] Furthermore, there are a total of 31 companies which collect data on over 25% of news pages. These findings demonstrate that a handful of companies have tremendous power to monitor the visitors of a wide spectrum of news websites.

The concentration of data flows is a feature which is more pronounced in the news population than on other popular web sites. In the reference population, only Google receives data on more than half of sites and only three companies receive data on more than a quarter. The upshot of this finding is that while Google has tremendous power on popular sites in general, the other companies' power is significantly more concentrated in the news sector.

[6]It should be noted that for Amazon, only 15.16% of pages use Amazon's advertising system, the majority of Amazon-directed requests are for possibly non-advertising content hosted on Amazon Web Services.

Table 6.3: Top Ten Data Recipients

News		**Reference**	
Company	Percent of Pages	Company	Percent of Pages
Google	94.99	Google	83.35
Facebook	60.45	Facebook	39.48
Oracle	55.04	Amazon	27.32
Amazon	51.39	AppNexus	21.57
comScore	48.67	Aol	20.09
The Trade Desk	46.06	Twitter	16.93
AppNexus	45.17	Yahoo	16.72
Acxiom	43.92	Adobe	15.74
Lotame	42.05	Oracle	15.18
Twitter	40.02	The Trade Desk	14.46

News websites likely have a high centralization of data flows for two reasons: first, the embrace of OBA results in unchecked flows of user data, and second, publishers have outsourced the hosting of their content to outside parties. In regards to revenue, Pew's 2015 State of News Report revealed that Google, Facebook, Microsoft, Yahoo and Aol were responsible for "61% of total domestic digital ad revenue in 2014, $30.9 billion out of a total $50.7 billion", with Google alone accounting for 38% of digital revenue (Mitchell and Rosenstiel, 2015). It is therefore no surprise that Google is found on 95% of sites. Furthermore, Pew found that digital advertising on news websites is dominated by "display ads such as banners

or video" as opposed to "search ads" (Mitchell and Rosenstiel, 2015). These types of ads rely on behavioral data for targeting, which is only possible when data is collected from a large range of sites. Thus, an independent publisher is simply unable to provide targeting on par with the large OBA companies. The inability to compete with massive OBA companies makes publishers dependent on large firms for their very survival.

In addition to revenue, news websites also rely on third-parties to deliver website content. Using `webXray`, it is possible to explore the dynamics of content hosting by analyzing the literal *volume* of data transferred on a per-company basis. Simply put, each HTML file, image, and piece of Javascript code is embodied in a variable length of ones and zeros which must be transferred across the Internet. Transferring this data takes physical material resources (e.g. electricity, computer hardware, air conditioning, server racks, etc.) which are comparable to what was formerly needed to deliver physical newspapers (e.g. paper, printing presses, delivery persons, news kiosks, etc.). Rather than host independent data centers, most publishers now outsource the hosting of advertisements and page content to a handful outside parties.

As with the flows of data to companies, the volume of data by company reveals a much more centralized system than many could imagine. In both the news and reference populations Google is responsible for hosting and transferring roughly 1/5th of all web traffic (21.77% for news, 18.42% for reference). This means that not only

does Google surreptitiously collect information on users and control revenue, they support so much of the hosting infrastructure that if they were to stop providing service, most websites would cease to be fully functional due to missing page components. This gives Google a tremendous amount of hidden power over the web which goes far beyond search dominance - even if publishers were to reject a new Google policy and decide to end their relationship, they would be hard-pressed to host the requisite page content themselves. Not only can Google de-list a website from search results at will, they may also be able effectively pull the plug on a website if they wish.

Beyond Google being fairly even between news and reference sites, news websites exhibit greater centralization of data flows across the top ten companies as seen in Table 6.4. Furthermore, 83.41% of all data on news websites comes from a third-party server, compared to to 57.81% for reference sites, which indicates that news websites have lower capacity than average to independently host their own content. This is partially explained by the larger volume of advertisement which are frequently delivered from OBA companies rather than publishers.

Table 6.4: Volume of Data Transfer by Company

News		**Reference**	
Company	Percent of All Data	Company	Percent of All Data
Google	21.77	Google	18.42
Facebook	7.55	Facebook	5.81
TownNews	4.68	Amazon	2.29
Twitter	2.44	Twitter	2.14
Amazon	2.44	Oracle	0.95
Taboola	2.39	Automattic	0.77
Oracle	2.33	Yahoo	0.65
Tout	1.79	Alibaba	0.46
Inform	1.67	Taboola	0.37
Krux Digital	1.38	StackPath	0.37

An additional risk to publisher independence is that reliance on OBA makes their web pages more expensive to view and slower to load than other websites. In regards to expense, users on metered data plans (such as mobile) must pay their Internet service providers for the data they download, and larger pages cost more money to view. On average, news websites are 3.95 megabytes in total size compared to 1.57 megabytes for a page in the reference population, a nearly 40% difference.[7] On a limited data plan, a user may view fewer news websites than other types of sites for the same money. Furthermore, when considering the time needed

[7]Note these measures exclude Flash video content which would further skew the numbers.

to download a page, it is found that news websites take 12.89 seconds on average to download compared to 9.23 seconds for reference web sites, meaning that readers of news sites must wait 30% longer to get content even if they have unlimited plans and are unconcerned about cost.

The threat to publisher independence from large and slow web pages is not that users will stop being interested in news, it is that other companies will take publisher's articles and deliver them in a streamlined environment. In fact, the two biggest collectors of data on news websites, Google and Facebook, offer services by which news articles are delivered quickly in an aggregated format. This gives these companies control of what articles get presented to users, eliminating the independence of news editors, and reduces the likelihood users will browse many pages from a single media outlet. The great irony is that OBA degrades the experience on news websites while simultaneously siphoning user data in a way that allows OBA companies like Google and Facebook to steal visitors from publishers.

While it is clear that OBA represents a diminishment of publisher independence on the web, independence could be regained by adopting new revenue and hosting models. However, a much greater risk to the viability of news outlets is the loss of public trust, which is difficult, if not possible to regain.

Trust: Analysis of Policy Disclosure and Readability

As noted above, journalistic trust is deeply rooted in full and fair reporting of the facts. When it comes to data flows on newspaper websites such "reporting" happens in the privacy policy. In order to determine if a given web pages' privacy policies are both accurate and clear they have been analyzed across two dimensions: accuracy and clarity. In terms of accuracy, for each page with a privacy policy the entities receiving data on the page are searched for in the body of the policy text using `policyXray` in order to verify disclosure. In regards to clarity, widely accepted readability metrics are used to determine if policies are written in such a way as they may be understood by an educated reader.

In order for a news website to be fully transparent with users about how their data is collected and used all of the companies which collect user data must *at least* be mentioned in the policy text. However, policies for news websites fail to disclosure the vast majority of data flows taking place and may be considered at best incomplete, and at worst intentionally misleading. Only 3.58% of data flows are disclosed on the 140,340 news pages with a corresponding privacy policy. In the reference population, 6.03% of data flows are disclosed across 26,758 sites which had policies. Once again, both news websites and reference sites fare poorly, with news websites continuing to engage in practices which are worse than the norm.

Although the rate of disclosure is *universally* incomplete across pages, it is not *uniformly* incomplete across the various companies which have been detected. Many

companies are never mentioned in *any* privacy policies, meaning even users who make the effort to read policies will remain ignorant of data flows. In the news population, 126 companies were found collecting data, of these 87 (69.04%) are never mentioned. In the reference population, 169 companies were found collecting data, of these 61 (36.09%) were never mentioned. Those who read privacy policies of news websites are therefore nearly twice as likely to be tracked by a company whom they have no means of discovering.

Table 6.5: Disclosure for Top Ten Data Recipients (News)

Company	**Percent of News Pages Tracked**	**Percent News Disclosed**
Google	94.99	28.06
Facebook	60.45	33.91
Oracle	55.04	4.14
Amazon	51.39	0.08
comScore	48.67	2.91
The Trade Desk	46.06	0.01
AppNexus	45.17	2.32
Acxiom	43.92	0.06
Lotame	42.05	0.09
Twitter	40.02	16.08

While many companies are never mentioned, those that are disclosed are done so at vastly different rates. Table 6.5 demonstrates that among the companies which receive the greatest percentage of user data on news websites, disclosure ranges from

Facebook being disclosed roughly one-third of times down to Amazon and Lotame who are mentioned in fewer than 1% of applicable policies.

Table 6.6 contains a ranking of the top ten companies according to their rates of disclosure in news website policies. Of these companies, only Facebook, DataXu, and Google are mentioned in more than a quarter of cases. By ninth place disclosure rates hit the single digits (6.04% for Adobe) and the distribution of disclosure follows a "long tail" where the majority of companies have rates of disclosure which are slightly above zero (for example, the company Centro is disclosed in 0.01% of cases).

Table 6.6: Top Disclosures of Tracking

Company	**Percent Policies with Disclosure**
Facebook	34.31
DataXu	30.31
Google	28.10
NetIQ	23.04
Twitter	16.08
Turn	15.89
Collective	12.90
Share This	10.07
Adobe	06.04
Yahoo!	05.95

An additional facet of disclosure is if the privacy policy specifies the site's response to the "Do Not Track" standard. As noted in Chapter Four, "Do Not Track"

is a setting available in all major web browsers which tells websites and embedded elements the user does not wish to be tracked. There is no enforcement mechanism and it is up to the party receiving the request to honor it and to be clear with users what the policy is. In the news population only 15.93% of policies mention the standard, which is actually better than the reference population where 10.41% mention DNT. It should be noted that merely mentioning the standard does not mean it is respected by the site or the trackers on the pages, and this is a very low bar for disclosure.

From the above it is clear that the policies are far from accurate and demonstrably incomplete. Despite this, it is still possible they would contain other valuable information about the data retention policies of the news outlet. Yet for such information to be useful, it would need to be clearly stated. Unfortunately for potential news site privacy policy readers, the average reading difficultly score for all 2,881 news policies is in the college-level difficulty range (38.43 on the Flesch Reading-Ease scale and a 13.87 Flesch-Kinkaid grade level score). Similar findings were evident in the reference population where policies were also in the college-level difficulty range (38.52 on the Flesch Reading-Ease scale and a 13.57 Flesch-Kinkaid grade level score). While a non-trivial portion of users may have college degrees, privacy policies are incredibly difficult to read, indicating that in the realm of journalistic clarity and truthfulness, they fall short of established norms.

The impact of low disclosure rates for data flows is that even if users make the

significant effort to read privacy policies, they are highly unlikely to learn about the companies which receive their data. If editors and reporters decided to leave out key facts from news stories they would quickly lose the trust of readers, if they failed to report 97% of relevant facts they would quickly go out of business. Viewed from a perspective of transparency and trustworthiness, rates of disclosure of data transfers clearly shows how the spread of OBA undermines journalistic norms. Beyond not trusting news websites to disclosure data flows, users should not trust them to keep their computers safe from criminals either.

Trust: Analysis of Malware Exposure

In addition to a lack of accurate disclosure in privacy policies being a reason not to fully trust a news website, the degree to which a given site is associated with fraudulent and criminal malware schemes provides an even stronger indication. OBA allows advertisers to send Javascript code to web browsers of those viewing ads. This code is poorly vetted and criminals have discovered that buying advertisements is a profitable and effective means of delivering malicious software in what has been termed "malvertising" schemes. Most troubling of all, criminals may target users for infection in the same way advertisers target them for discounts.

The results of the `webXray` analyses for both the news and reference populations are augmented with data from the VirusTotal malware database. The VirusTotal database is a free service provided by Google, and is an aggregation of "different

antivirus engines, website scanners, file and URL analysis tools and user contributions". [8] The VirusTotal API queries roughly 60 anti-virus sources in order to report if a specified domain has been marked as hosting malware. In order to provide a strong basis to draw conclusions, only domains that are marked by more than one source as malware are considered to be harmful.

For both the news and reference populations, two measures were taken: first, the number of pages that themselves had been previously marked as malware hosts, and second, the number of pages which initiate HTTP requests to third-party domains which have previously been identified as a malware host. As can be seen in Table 6.7, 2.8% of news and 6.77% of reference sites have been themselves marked as hosting malware. In this case of news sites the greatest likelihood for this happening is the site having been compromised (or "hacked"), in the case of the reference population, the greatest reason for this happening is the Alexa list may include malware sites. Regardless of cause, it is alarming that several hundred news websites have been marked as delivering malware.

Table 6.7: Domain Reputation and Malware Exposure

Population	Pages Identified as Malware	% Pages w/Third-Party Domain Linked to Malware
News	2.88	80.82
Reference	6.77	67.87

[8] https://www.virustotal.com/en/about/

Providing an insight into the problem of malvertising are the rates at which pages include requests to domains which have previously hosted malware. Table 6.7 contains the results of checking 6,018 third-party domains in the news population and 40,556 domains in the reference population for hosting malware with the VirusTotal API. 80.82% of pages in the news population link users to domains which have previously hosted malware, while 67.87% of pages in the reference population have. Once again, news websites present fewer reasons for users to trust them - in this case by exposing them to criminals due to their embrace of OBA.

On a darkly ironic note, 53 of the domains linked to malware found on news websites contained the word "safe" in them. Examples are "updatestogetforfeesafe.space", "internetssafety.com", and "24online5678safesystems.site". Such domains may fool many users, but for the professionals responsible for coding and maintaining websites, they should represent red flags which demand intervention. The failure to spot such obvious scam domains is clear negligence, and represents yet another reasons not to trust news websites.

While VirusTotal provides an excellent means to scan large numbers of domains, it does not mark any of Google's advertising domains as malware. This is despite the fact that in March 2015, security firm Malwarebytes published research demonstrating that the websites of MSN, The New York Times, The BBC, Aol, and others were delivering malware payloads to visitors over advertisements delivered by domains owned by Google, AppNexus, Aol, and Rubicon. As noted above, Google alone is

found on 95% of sites, indicating the problem of malvertising may be even worse than VirusTotal data suggests. Not only does the use of OBA make news websites poor guardians of user privacy, it helps expose users to criminals in such a way that any sense of trustworthiness is eliminated entirely. However, when it comes to the deliberative function of the press, malware is among the least concerning aspects of OBA.

Chilling Effects and Potential Exposure to State Surveillance

As noted above, one of the primary roles of the press is to facilitate democratic deliberation by citizens. One of the means by which such deliberation may be compromised is through the "chilling effects" of surveillance. Thus, if news websites facilitate state surveillance they ultimately put their ability to promote deliberation at stake. This study reveals that OBA opens many avenues for state surveillance.

Behavioral tracking on news websites potentially exposes users to two forms of unwarranted state surveillance. In the first, major Internet companies may either be compromised or forced to disclose users' web browsing information to authorities. In the second, companies which sell data directly to the government may decide to sell web browsing information collected from news media websites.

In 2013, former U.S. National Security Agency (N.S.A.) contractor Edward Snowden leaked a massive trove of internal documents to the Guardian and The Washington Post (both of whom won Pulitzer prizes for reporting on the contents

of the leaks).[9] One of the first reports to emerge from the Snowden leaks revolved around a project named "PRISM" (Lee, 2013). According to an internal presentation, PRISM allowed the N.S.A. to collect information "directly from the servers" of Microsoft, Yahoo, Google, Facebook, PalTalk, Aol, Skype, YouTube, and Apple (Ball, 2013). The companies, however, denied these allegations and several years later the exact nature of PRISM remains a mystery. Regardless of details, it is undeniable that if one were to gain access the databases of the companies mentioned in the PRISM slides a virtual treasure trove of web browsing information would be found.

By using OBA, news websites funnel user data to the very databases most sought after by intelligence agencies. Based on corporate data flow findings from the above section, it is possible to determine if PRISM-linked companies Google, Facebook, Aol, Yahoo!, and Microsoft may monitor a given web page. It is found that not only do these companies all collect data from the websites of news organizations, they do so at far higher rates than seen in the reference population (see Table 6.8). This finding further underscores how the loss of independence by publishers results in user data being centralized under the authority of entities outside the control of publishers; in this case, the loss of independence may result in state surveillance.

[9]https://www.theguardian.com/media/2014/apr/14/guardian-washington-post-pulitzer-nsa-revelations

Table 6.8: Exposure to PRISM Companies

Company	News	Reference
Google	94.99	83.35
Facebook	60.45	39.48
Aol	39.78	20.09
Yahoo!	36.10	16.72
Microsoft	17.84	8.51

While the PRISM slides raised more questions than answers, other slides Snowden leaked contained far more concrete details on how the N.S.A. has used data from online behavioral advertising for spying purposes. In a 2013 article, "NSA uses Google cookies to pinpoint targets for hacking", *The Washington Post* detailed how the N.S.A. was repurposing Google cookies for their own ends. According to former Deputy U.S. Chief Technology Officer Ed Felten, the slides demonstrate "a link between the sort of tracking that's done by Web sites for analytics and advertising and NSA exploitation activities" (Soltani et al., 2013). Another article in *The Guardian* revealed a specific Google cookie, "Doubleclick ID", was used in efforts to spy on users of the Tor anonymity service (National Security Agency, 2013).

Table 6.9: Exposure to Google Cookies

Population	Any Google Cookie	Doubleclick ID
News	76.01	62.85
Reference	49.94	35.71

An investigation of the cookies set when loading the news and reference websites shows that 76.01% of news and 49.94% of reference pages include a Google cookie of some sort. In regards to the specific "Doubleclick ID" cookie, 62.85% of news and 35.71% of reference sites utilize cookies known to be leveraged by the N.S.A. As with the findings above, news sites consistently demonstrate significantly greater threats to privacy than the reference population.

Beyond data found in the Snowden archives, prior research and reporting has documented how so-called "data brokers" may sell personal information to military and law enforcement organizations. Per Table 6.3, one of the largest data brokers, Acxiom, was found on 43.92% of news media pages analyzed. According to an internal email regarding the now-defunct U.S. Department of Defense "Total Information Awareness" project, a military official discussed obtaining Acxiom's data with the company's Chief Privacy Officer over a decade ago (Hoofnagle, 2003). While Total Information Awareness was formally disbanded, the internal emails made it quite clear the military had a strong interest in Acxiom's data.

Several years after Total Information Awareness was shuttered, a 2009 Wired

report revealed that the FBI's National Security Branch Analysis Center (NSAC) possessed "nearly 200 million records transferred from private data brokers such Accurint, Acxiom and Choicepoint" (Singel, 2008). Based solely on the findings of the research presented herein, it is impossible to determine if the data sold by Acxiom to the F.B.I. includes web browsing histories. However, it is quite clear there is certainly data to be sold.

Another company receiving large volumes of user data from news websites is Oracle whose "AddThis" subsidiary captures user data on 55.04% of pages. AddThis' main product is a social media sharing tool (widget) which generates no direct revenue for publishers and provides functionality which is trivial to implement. Despite the relative uselessness of the AddThis widget for users, significant volumes of their data are harvest by the widget. The data itself is quite valuable and explains why Oracle paid $200M to acquire AddThis in 2016.[10] According to Oracle's press release, the AddThis acquisition gives the company the ability to collect online behavioral data in order to "enable[s] understanding of consumer behavior across all media channels".[11]

Of particular importance to the subject of chilling effects, Oracle has deep ties to the U.S. government. According to an article in The Guardian, CEO Larry Ellison took the name for his company from a CIA project he worked on codenamed

[10]https://www.crunchbase.com/acquisition/c1da0a6668ddf5b33c9360496a7d5b43

[11]https://www.oracle.com/corporate/acquisitions/addthis/index.html

"Oracle".[12] Furthermore, among Oracle's many clients are those of the "Public Sector for Defense" division which "delivers a comprehensive set of applications and commercial off-the-shelf solutions that help military organizations improve efficiency, mission preparation, and execution".[13] Likewise, Oracle's "Immigration and Border Control" services assist with "managing the tracking of individuals within national boundaries".[14] It is not possible to determine if the data AddThis collects from news websites is made available to Oracle's government clients, but an evaluation of AddThis' privacy policy also reveals there are no guarantees it is not shared.

The nature of chilling effects is that the mere impression of surveillance may limit debate and inquiry, which are the core functions of the press in a democratic society. The above links between OBA and state surveillance are *not* definitive proof that the N.S.A., F.B.I., U.S. military, or other such entities spy on visitors to news websites by leveraging mechanisms and data collected by OBA firms. However, the possibility this may happen is strong, and the possibility *alone* is enough to chill speech. The case of AddThis, which provides scant benefit to publishers, but significant risks for personal privacy, underscore the degree to which publisher ignorance may be driving the trend. Furthermore, it is precisely the move to OBA from traditional advertising which has opened up this new venue of surveillance.

[12]https://www.theguardian.com/g2/story/0,3604,215072,00.html

[13]https://www.oracle.com/industries/public-sector/defense.html

[14]http://www.oracle.com/us/industries/public-sector/046927.html

Conclusion

Through an in-depth investigation of how online behavioral advertising (OBA) functions on news websites, this chapter has demonstrated that OBA is in many way anathema to core democratic norms represented by a free, independent, and trustworthy press which stands as a bulwark of freedom. OBA is rapidly transforming publishers who once had the power and independence to keep government in check into branded content providers who help create audiences which are auctioned off by OBA companies. Likewise, the embrace of the unsavory and inherently deceptive aspects of OBA has undermined the trustworthiness of the press, putting them in league with both unethical businesses and criminals. Finally, by facilitating state surveillance, OBA places a potential chill on the ability of the press to facilitate public deliberation. In sum, OBA weakens the democratic potential of the press, undermines essential social norms, and centralizes power into fewer hands where it may be abused.

In the previous chapters on fair information practices and medical privacy, the failure of regulators was clear, but they were at least visible in the debates. When it comes to the ways in which OBA undermines press freedoms, regulators are nowhere in sight. Users, however, are taking protection into their own hands, and punishing news outlets for their use of OBA. This punishment comes in the form of using software tools to disrupt the functioning of OBA, and deny publishers advertising revenue.

According to multi-nation study by the Reuters Institute for the Study of Journalism at Oxford University, in some countries as many as 38% of users are now using ad-blocking software to keep intrusive and harmful advertisements off of their computers (Newman et al., 2016). Blocking ads denies publishers revenue and the "vast majority of those who have ever downloaded a blocker are using them regularly, suggesting that once downloaded people rarely go back" (Newman et al., 2016). In the majority of countries surveyed over half of users of ad-blockers reported doing so due to a dislike of OBA-specific "ads that follow [users] around from one site to another" (Newman et al., 2016).

Due to the fact that there is a long history of newspapers competing with each other for "scoops" and advertisers, news media outlets have failed to recognize that their true competitors are not other media outlets, but the OBA industry. From a revenue standpoint, companies like Facebook and Google do not need to worry if an advertisement is run on the page of a reputable news outlet which provides a meaningful contribution to society, or a scam webpage peddling falsehoods: click revenue is made either way.

However, what *does* matter is the speed by which additional ads may be shown to users, thus the goal for these companies is to speed up delivery of content by taking full control of the editorial and hosting functions of websites in order to manipulate users to keep clicking on articles and advertisements. Facebook and Google may claim to value the press, but from a business standpoint news outlets

are merely another form of media to be "disrupted" and subsumed into a centralized revenue system which exploits users for commercial gain.

To the degree that there has been industry coordination on the topic of OBA, it has been woefully misguided. Initiatives such as "The Coalition for Better Ads"[15], which is supported by the trade group "News Media Alliance", are seeking to develop standards that "that reflect consumer advertising preferences".[16] However, these standards are mostly aimed towards user-interface concerns such as annoying full-screen ads. Privacy concerns are barely mentioned, acknowledgement of malvertising is nowhere to be found, and topics such as chilling effects are miles away from industry concerns. However, if the online news media industry wishes to continuing fulfilling a democratic function, they need to quickly realize that the problem is much bigger than users being annoyed.

A free and independent press is the ultimate defender of democratic freedoms. Online news websites, however, appear unable to defend themselves from the predatory practices of the OBA industry. Thus, it is up to regulators to step in and bring fundamental changes which cannot only improve user privacy, but save the news media from being demoted to content providers for the likes of Facebook and Google. The final chapter of this dissertation proposes a radically new approach to regulation which may assist in this vital project.

[15]https://www.betterads.org/about/

[16]https://www.newsmediaalliance.org/release-cba-better-ads-standards/

Chapter 7

Conclusion: Surveillance as a Regulatory Model

The methods and normative frameworks used in this dissertation have revealed a picture of a powerful online behavioral advertising industry pursuing goals which are at sharp odds with the public interest. The technological systems deployed by this industry have created a ubiquitous system of surveillance which implicates nearly all facets of social life and empowers the state to conduct mass surveillance of citizens. These are legitimate causes for serious concern, and given the extent of the problems, some may adopt a stance of resignation and abandon the idea of privacy as a foundational social value. However, there are still avenues available which may produce regulatory outcomes which ensure respect for established privacy norms and promote social justice.

This concluding chapter of the dissertation revisits and summarizes the ways in which the current self-regulatory paradigm is failing to protect essential privacy norms on the web as well as why purely technical controls have been insufficient to protect privacy. Finally, a radical new approach to regulating online data flows will be proposed: mass surveillance of the companies which track users' behavior.

Current Approaches to Controlling OBA Do Not Work

The current regulatory approach to online tracking in the United States relies on industry self-regulatory guidelines which are largely focused on the paradigm of "notice and choice". This paradigm is a conceptually flawed interpretation of the rights-based Fair Information Practice Principles (FIPPs), is practiced in a manner which undermines social norms related to health and the democratic function of the press, and if "notice and choice" were to be practiced in a meaningful way, it would make the web unusable.

As noted in Chapter Four, the FIPPs were developed in 1973 as a normatively-grounded "bill of rights" designed to protect "private individuals (data subjects) against large government and private sector institutional actors" (Nissenbaum, 2004). The FIPPs were originally constituted by five broad principles aimed at governing procedures for data processing in a way which embodied specific "norma-

tive political views" (Rotenberg, 2001). These views were grounded in a philosophy which regarded privacy as an essential human right. At the time of their writing, the FIPPs represented a prescient view on emerging challenges to privacy in the age of digitization, and proved to be highly influential. FIPPs were broadened over time to include additional principles, and for the past forty years FIPPs have formed the basis for privacy regulations around the globe. The wide adoption of FIPPs in national data regulations provides strong evidence that a rights-based approach to privacy has near-universal support.

However, in the realm of online privacy, the U.S. Federal Trade Commission eschewed the broad rights-based consensus on FIPPs in favor of the "notice and choice" self-regulatory paradigm. This paradigm reduces the essential rights approach embodied in FIPPs to a system where data subjects are framed as *consumers* rather than *citizens* or *natural persons*. The impact of this seemingly minor change in framing is that it transfers "the protection of privacy from the legal realm, and from an emphasis on the articulation of rights and responsibilities, to the marketplace, where consumers [are] forced to pay for what the law could otherwise provide" (Rotenberg, 2001). Rather than a system designed to protect privacy, notice and choice is a "creation of the U.S. marketing industry" which promotes consumerism and profit (Rotenberg, 2001). Most egregious, the "choice" offered by the online behavioral advertising industry is not a choice in regards to data collection, it is a choice to see "tailored ads" instead of generic ads. Thus, the *actual* self-regulatory

policies followed by industry today hold that ***users have no choice to control data collection***. The elimination of the right to control data collection has broad social consequences.

Ubiquitous non-consensual data collection on the web leads to harms in several areas of social life, two of which are medical privacy and the role of the press in a democratic society. In regards to the former, there is a long history of medical privacy norms which continues to this day. Patient information is protected by professional guidelines and numerous laws because the exposure of patient information may lead to shame and discrimination. Likewise, if one is concerned about these harms, she or he may avoid seeking medical care which may increase medical costs, foster the spread of disease, and raise mortality rates. Chapter Five demonstrated that these norms are greatly undermined by the large amount of tracking happening on medical-related websites. Such tracking is not limited to commercial websites and includes non-profit and government websites which employ information-leaking social media and traffic analytics code. Despite the huge volume of tracking taking place on medical websites, the practice is virtually unregulated in the United States and self-regulatory guidelines offer scant, if any, protection for users.

In order to fulfill its democratic purpose, the press needs to be independent, trusted by citizens, and able to facilitate public deliberation on matters of governance. However, OBA undermines all of these functions. The emergence of a handful of powerful advertising networks has centralized the flow of user data,

and by extension, the mechanisms for revenue generation and content delivery. These trends have begun to transform independent media outlets into mere content providers for companies like Facebook and Google. Furthermore, the lack of transparency in data collection practices and risks brought about by malvertising serve as a basis to seriously undermine the trust citizens may place in the press. Finally, the ubiquitous surveillance carried out by online advertising networks has provided law enforcement and intelligence agencies with the ability to monitor citizens as they read the news. The mechanisms of OBA invert the role of the press from a protector of freedoms into a facilitator of state power. The most profound impact of OBA on the press is that it may chill speech, and reduce the ability of the press to foster democratic deliberation. While the connections between OBA and the social utility of the press are subtle, the impact is profound.

Assuming that the normative issues detailed above are of concern to the OBA industry (and there is scant evidence they are), it may be argued that refinements to the "notice and choice" system could be made. However, considering the volume of tracking, it is utterly infeasible to give users sufficient notification of all of the data transfers taking place. For example, the EU requires notifications from webpages using cookies. Such notices often take up a portion of the top of the page and annoy users by requiring them to click a button which expresses consent. Despite this, the notices are usually vague and fail to disclose the entities receiving user data. If EU-stye cookie notices were to actually include relevant details for each cookie,

the contents of a given page would be obscured by numerous notifications. For the Alexa top one million sites, *eleven* cookie notices would be required on average, and on news websites *twenty-seven* notifications would be required. Assuming a user clicks one notification per second, it would take nearly half a minute to consent to viewing a given page. Even worse, these examples include only cookies. If users were also notified of each third-party domain receiving their data, it would take well over a minute of clicking consent buttons to view a page on an average news website. It is amply clear that the volume of tracking makes meaningful notification impossible without destroying the usability of the web.

On both theoretical and practical levels, it is clear that self-regulation in the OBA sector is failing to respect and protect vital social norms. Due to this failure, advocates of user privacy have developed a range of technical mechanisms to protect privacy. Unfortunately, this approach is flawed as well.

As the issues raised by OBA are of a technical nature, at first glance they invite technical solutions. Indeed, privacy-protective technologies such as browser "add-ons" have proven effective at preventing certain types of behavioral tracking (Mayer and Mitchell, 2012; Roesner et al., 2012), and are enjoying increased adoption (Newman et al., 2016). However, there are several limitations of purely technical ad-blocking approaches: they foster a cat-and-mouse dynamic between researchers and OBA firms, technical measures often unduly place the burden of employing privacy protections on users, and blocking ads may deprive already-

vulnerable publications of sorely needed revenue.

A major reason that purely technical solutions have not been broadly effective is that whenever an effective anti-tracking methodology is demonstrated by researchers it is swiftly met with technical counter-measures by industry which keep user data flowing. This has produced an adversarial dynamic between companies and researchers in which the protection of user data has become a high-stakes game. For example, the rise of cookie-blocking resulted in the adoption of fingerprinting. The use of ad-blockers has led to an emergence of ad-blocker-blockers which prevent users from viewing content. Apple's efforts at limiting tracking in the Safari browser led Google to pursue deceptive means to overcome the mechanism, resulting in a $17 million FTC settlement (Fung, 2013). Despite the threat of fines, industry players with deep pockets continue to hire top-tier computer scientists to develop new ways of tracking users and evading control. Researchers acting in the public interest simply do not have the resources needed to keep up with their adversaries. Although there should be a continued effort to develop technical counter-measures, the game being played is unending and has not slowed the growth of OBA to any meaningful extent.

Furthermore, purely technical solutions are problematic as they require a relatively high level of knowledge and technical expertise on the part of the user. The user must first understand the complex nature of information flows online in order to seek out technical remedies. Next, the user must be proficient enough to

install and configure the appropriate browser add-ons. This may seem trivial for the well-educated, but many who use the Internet have little education or training in computing. Furthermore, users are spending increasing time in mobile browsers, where it is significantly more difficult, if not impossible, to install browser add-ons. Ultimately, technical counter-measures have the unintended side-effect of shifting responsibility for privacy protection from companies and regulators to the subjects of behavioral tracking. The shift of responsibility to users is the same flaw which undermines the essential legitimacy of the "notice and choice" framework in the first place, ad-blocking simply reinforces this dynamic.

Finally, technical countermeasures often result in the elimination of advertising revenue which disproportionally affects online publishers who are "almost entirely display ad supported" (Budak et al., 2016). Google, Facebook, and other companies make billions of dollars in online advertising and have diverse product portfolios which span search, social media, mobile communications platforms, and web display advertising. This diversity means that while a specific revenue line may be temporarily threatened by a new blocking technique, large companies are structurally resilient to such attacks. For these companies, short-term financial losses due to blocking are not mortal wounds. In contrast, small-time publishers may rely solely on display advertising and operate on incredibly tight margins. Losing a small amount of revenue to ad-blocking may put such publishers out of business. The impact of this trend is that users will get even more of their content from Facebook

and Google as independent outlets disappear. In a sadly ironic twist, ad-blocking may strengthen the giants of OBA rather than hurting them.

Purely technological fixes to online privacy issues are insufficient as they are at-best temporal in the face of industry counter-moves, place an undue burden on users, and harm publishers. However, technological approaches may be leveraged to provide assistance to *regulators* rather than users.

Surveillance as a Regulatory Model

Bruce Schneier has noted that "surveillance is the business model of the internet" (Schneier, 2013). Given that the current self-regulatory approach to online privacy on the web is failing, it is necessary to ask what approach can control surveillance-based businesses? The answer is counter-intuitive, but addresses the underlying information asymmetry which empowers online behavioral advertisers at the expense of users. In the same way users live under the gaze of the OBA industry, regulators should engage in constant and ubiquitous surveillance of the practices of the OBA sector. I terms this approach "surveillance as a regulatory model".

Using the tools developed for this dissertation (`webXray` and `policyXray`), it is now possible to constantly monitor the activities of companies tracking users on the web and provide regulators with daily and historical reports on the state of tracking. Such an approach has several benefits: first, it streamlines the current process for documenting and punishing privacy infractions in specific contexts; second, au-

tomated code review systems may detect new deceptive practices as they emerge and raise alarms; third, historical records may facilitate retroactive punishments; fourth, automation makes it feasible to monitor large numbers of companies simultaneously; and finally, surveillance as a regulatory model may produce a form of chilling effects by which website operators and OBA companies become significantly more cautious in their approach to user tracking.

When companies get fined for privacy violations on the web it is often through a convoluted and inefficient process which combines academic research, publicity, and regulator action. Quite often, academic researchers will discover a new tracking technique in the wild and either publish findings online or at academic computer science conferences. These findings often get publicized in popular media as they tend to impact the majority of web users, raise issues of legitimate public concern, and feed into extant negative opinions of behavioral advertisers held by the public. Publicity often results in public deliberation which places pressure on regulators to levy fines against the most powerful companies in order to set an example to the wider industry. This process is not a universal formula, but it is a familiar pattern and demonstrates a reactive rather than proactive approach to regulation.

A much better approach is to separate research findings from the academic publishing cycle and allow regulators to continually monitor websites in order to issue fines quickly. Given the maturity and stability of `webXray` and related tools, regulatory agencies around the world should devote modest budgets for either the

operation of web privacy measurement systems or the licensing of centrally collected data (e.g. via a API or reporting framework). Large scale scans of data flows may be continually performed on popular websites, websites located in or serving specific geographic jurisdictions, websites related to special social contexts such as health or sexuality, and specific types of content such as news. The same economic forces which facilitate companies monitoring users, affordable cloud computing and cheap data storage, likewise make mass surveillance of companies engaged in behavioral tracking entirely feasible. Rather than reading about new privacy violations on the web in newspapers, regulators may instead generate positive publicity for their agencies and win public accolades by catching companies when they launch new tracking mechanisms.

Currently, the primary area of innovation in regards to tracking users is in the use of constantly evolving Javascript fingerprinting techniques. To determine if, and when, such techniques violate regulations such as deceptive trade practices, all Javascript deployed by companies in the OBA sector may be downloaded and stored. This code could then be analyzed in order to determine if it uses previously known deceptive fingerprinting techniques or exhibits characteristics such as specific function calls which indicate fingerprinting. In cases where an automated approach surfaces code of potential consequence, it may be evaluated by experts. If this process confirms an objectionable practice, companies should be contacted, fines may be levied, and the public should be informed.

While an alerting system may reveal new practices in short order, it should not be assumed that they will always be detected. Academic research will continue to proceed independently of regulatory actions and discover new tracking methods. Thus, in cases where outside researchers find new practices which have been in use for a significant period of time, regulators may consult their comprehensive historical records in order to determine when such practices began. Fines then may be levied for both current and prior violations, as well as multiplied by the number of sites and users affected. This approach would ensure that bad behavior does not go unpunished even if it happened in the past.

While Google and Facebook are well known companies, control huge portions of the online advertising industry, and are the public face of industry wrongdoing and fines, they are far from alone. Prior chapters in this dissertation have revealed nearly two hundred companies collecting user data on the web, and many more exist beyond those currently identified by `webXray`. Oftentimes the actions of small companies are under-researched or ignored by the press, yet they may collect data on many millions of users. By surveilling large numbers of companies with tools like `webXray`, it may be possible to catalogue and identify a huge range of actors and provide regulators with a number of targets for enforcement actions. Likewise, with many companies under surveillance, it may be possible to evaluate the broader context of a particular tracking technique and select the company which is most vulnerable to legal pressure from a jurisdictional or social-contextual perspective.

Such an approach gives regulators a hand in picking the cases most likely to succeed, and sends a message to all companies that they are also being watched.

How will website operators and the OBA industry react to regulatory surveillance? One obvious way is they will begin to feel the type of chilling effects which limit speech in the public sphere, but instead of limiting the expression of controversial opinions they will limit the use of controversial code. This type of coding chilling effect would affect website operators and the OBA industry in distinct ways.

Ultimately, OBA companies cannot force website operators to use their code, and it is the responsibility of web masters to audit the code they include from third-parties. However, it is currently common practice for a developer to add code to a website without any oversight from compliance or regulatory staff, even in sectors such as medicine where the risks to user privacy are most pronounced. Given that they too are at risk of being fined by ever-watchful regulators, website operators would have a strong incentive to caution staff to review code before it is deployed. In order to catch mistakes or potential bad behavior by trusted third-parties, website operators could surveil their own websites in order to identify mistakes before regulators do, and, when appropriate, notify regulators if trust placed in a third-party is proven unfounded.

Online behavioral advertisers, having the most to lose, would quickly learn to conduct thorough reviews before deploying new tracking techniques. They would have an incentive to keep careful inventory of the content featured on the sites

they track lest they become entangled in sector-specific regulation related to health or other topics. The knowledge that all security practices were being monitored and recorded may result in an immediate and full embrace of transport encryption. Ubiquitous surveillance, paired with the threat of fines, has the potential to put a deep chill on the enthusiasm of industry for developing new and more invasive forms of tracking.

Most importantly, surveillance as a regulatory model has the potential to resolve the information asymmetry which at the core of the problem: private companies know far more about the lives of the public than the public know of company activities. Regulators equipped with comprehensive knowledge of all tracking taking place on the web may seek out areas where key social norms are violated and levy the appropriate fines to rein in objectionable behavior. Ultimately, regulators act on behalf of the public and the unique tools I developed for this dissertation may prove to be powerful weapons in the fight for privacy.

Future Work

There are a number of areas where academic research into the social effects of Online Behavioral Advertising (OBA) may be continued. First, the effects of "algorithmic" discrimination may be researched using experimental methods (see Sweeney (2013) for one of the best studies in this area). Second, qualitative research may investigate the perspectives of the software engineers who build OBA systems and whose voices

are currently missing in extant scholarship. Third, additional sectors and social contexts may investigate web privacy issues affecting individuals' sexual interests, financial concerns, and online political activities. Fourth, although `policyXray` is currently only used for a basic analysis of privacy policy content, more advanced natural language processing techniques could be used to draw findings from the same data. Fifth, the various economic incentives which drive OBA should receive more attention from researchers, particularly in regards to how services such as Google Analytics are used by non-profit websites. Finally, by using large volumes of longitudinal data, machine learning and predictive approaches may be used to generate new insights into the scope and evolution of OBA.

In addition to academic research, policy work must continue as well. Given the failure in the United States by regulators and congress to protect privacy, consumer groups may lobby lawmakers on behalf of the public as well as establish their own institutions for auditing and policing online privacy. Civil society groups may also put pressure on companies to adhere to accepted human rights norms related to privacy. Likewise, there are areas for consumer and human rights groups to work together, as is the case with the new partnership between Consumer Reports and the Ranking Digital Rights project. Furthermore, despite the fact that federal legislators have been negligent in their duties to protect privacy on the web, individual states have significant leverage in instituting new laws which would place additional limits on the industry, as is already seen in the realms of medical privacy and data

breach notifications. Finally, in the same way that professional groups spell out ethics guidelines for practitioners of law and medicine, the same groups could set standards for websites used in the profession. There is no single regulatory or policy approach which will "fix" privacy online, but a number of approaches pursued in tandem may have significant impact.

Bibliography

Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., Gürses, S., Piessens, F., and Preneel, B. (2013). Fpdetective: dusting the web for fingerprinters. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, pages 1129–1140. ACM.

Ackerman, M. S., Cranor, L. F., and Reagle, J. (1999). Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, pages 1–8. ACM.

Acquisti, A. and Gross, R. (2006). Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Privacy Enhancing Technologies*, pages 36–58. Springer.

Allen, A. L. (2008). Confidentiality: an expectation in health care. *Penn Guide to Bioethics*.

America, P. (2013). Chilling effects: NSA surveillance drives US writers to self-censor. *New York: PEN American Center*.

Apple (2013). Apple unveils ios 7 (retrieved 2016-04-06). *http://www.apple.com/pr/library/2013/06/10Apple-Unveils-iOS-7.html*.

Ball, J. (2013). NSA's PRISM surveillance program: how it works and what it can do. *The Guardian*, 8.

Barocas, S. and Nissenbaum, H. (2009). On notice: The trouble with notice and consent. In *Proceedings of the Engaging Data Forum: The First International Forum on the Application and Management of Personal Electronic Information*.

Barocas, S. and Nissenbaum, H. (2014). Big data's end run around procedural privacy protections. *Communications of the ACM*, 57(11):31–33.

Bennett, J., Bennett, J., and Strange, N. (2015). Introduction: the utopia of independent media: independence, working with freedom and working for free. *Media Independence: Working with Freedom or Working for Free*, pages 1–28.

Bishop, L., Holmes, B. J., Kelley, C. M., et al. (2005). National consumer health privacy survey 2005: Executive summary.

Black, J., Steele, B., and Barney, R. (1999). Doing ethics in journalism: A handbook with case studies.

Blanke, J. M. (2006). "robust notice" and "informed consent:" the keys to successful spyware legislation. *Columbia Science & Technology Law Review*, 7:2–16.

Boudana, S. (2011). A definition of journalistic objectivity as a performance. *Media, Culture & Society*, 33(3):385–398.

Budak, C., Goel, S., Rao, J., and Zervas, G. (2016). Understanding emerging threats to online advertising. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 561–578. ACM.

Castelluccia, C., Grumbach, S., Olejnik, L., et al. (2013). Data harvesting 2.0: from the visible to the invisible web. In *The Twelfth Workshop on the Economics of Information Security*.

Cate, F. H. (2006). The failure of fair information practice principles. In *Consumer Protection in the Age of the 'Information Economy'*. Routledge.

Christians, C. G. (2009). *Normative theories of the media: Journalism in democratic societies*, volume 117. University of Illinois Press.

Deuze, M. (2005). What is journalism? professional identity and ideology of journalists reconsidered. *Journalism*, 6(4):442–464.

Dickens, C. J. H. (1859). *A tale of two cities*, volume 1. Chapman and Hall.

Digital Advertising Alliance (2013). Interactive survey of u.s. adults. *https://www.aboutads.info/resource/image/Poll/Zogby_DAA_Poll.pdf*.

DuckDuckGo (2016). Duckduckgo sources (retrieved 2016-04-06). *https://duck.co/help/results/sources*.

Duhigg, C. (2012). How companies learn your secrets. *The New York Times*, 16.

Dwoskin, E. D. E. (2013). Data broker removes rape-victims list after journal inquiry. *Wall Street Journal*.

Eckersley, P. (2010). How unique is your web browser? In *Privacy Enhancing Technologies Symposium*, pages 1–18. Springer.

Englehardt, S. and Narayanan, A. (2016). Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 1388–1401. ACM.

Felten, E. W. and Schneider, M. A. (2000). Timing attacks on web privacy. In *Proceedings of the 7th ACM Conference on Computer and Communications Security*, pages 25–32. ACM.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and Berners-Lee, T. (1999). Hypertext transfer protocol – http/1.1.

FourthParty (2013). Fourthparty is an open-source platform for measuring dynamic web content. *http://fourthparty.info/*.

Fox, S. and Duggan, M. (2013). Health online 2013. *Pew Internet and American Life Project*.

Fung, B. (2013). Why states are the big winner in the 17 million google-safari settlement. *The Washington Post*.

Gandy, O. H. (2003). Public opinion surveys and the formation of privacy policy. *Journal of Social Issues*, 59(2):283–299.

Gellman, R. (2016). Fair information practices: A basic history. *Self Published*.

Graber, M. A., D Alessandro, D. M., and Johnson-West, J. (2002). Reading level of privacy policies on internet health web sites. *Journal of Family Practice*, 51(7):642–642.

Habermas, J. (2006). Political communication in media society: Does democracy still enjoy an epistemic dimension? the impact of normative theory on empirical research. *Communication Theory*, 16(4):411–426.

Haggerty, K. D. and Ericson, R. V. (2000). The surveillant assemblage. *The British Journal of Sociology*, 51(4):605–622.

Himmelstein, D. U., Thorne, D., Warren, E., and Woolhandler, S. (2009). Medical bankruptcy in the United States, 2007: results of a national study. *The American Journal of Medicine*, 122(8):741–746.

Hoofnagle, C., Urban, J., and Li, S. (2012). Privacy and modern advertising: Most us internet users want "Do Not Track" to stop collection of data about their online activities. In *Amsterdam Privacy Conference*.

Hoofnagle, C. J. (2003). Big brother's little helpers: How choicepoint and other

commercial data brokers collect and package your data for law enforcement. *North Carolina Journal of International Law and Commercial Regulation*, 29:595.

Jackson, C., Bortz, A., Boneh, D., and Mitchell, J. C. (2006). Protecting browser state from web privacy attacks. In *Proceedings of the 15th International Conference on World Wide Web*, pages 737–744. ACM.

Jang, D., Jhala, R., Lerner, S., and Shacham, H. (2010). An empirical study of privacy-violating information flows in javascript web applications. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, pages 270–283. ACM.

Karppinen, K. and Moe, H. (2016). What we talk about when talk about "media independence". *Javnost-The Public*, 23(2):105–119.

Kovach, B. and Rosenstiel, T. (2007). *The elements of journalism: What newspeople should know and the public should expect*. Three Rivers Press (CA).

Krebs, B. (2013). Sources: Target investigating data breach. *http://krebsonsecurity.com/2013/12/sources-target-investigating-data-breach/*.

Krishnamurthy, B., Naryshkin, K., and Wills, C. (2011). Privacy leakage vs. protection measures: the growing disconnect. In *Web 2.0 Security and Privacy Workshop*.

Krishnamurthy, B. and Wills, C. (2009). Privacy diffusion on the web: a longitudinal

perspective. In *Proceedings of the 18th International Conference on World Wide Web*, pages 541–550. ACM.

Krishnamurthy, B. and Wills, C. E. (2006). Generating a privacy footprint on the internet. In *Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement*, pages 65–70. ACM.

Lécuyer, M., Ducoffe, G., Lan, F., Papancea, A., Petsios, T., Spahn, R., Chaintreau, A., and Geambasu, R. (2014). Xray: Enhancing the web's transparency with differential correlation. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 49–64, San Diego, CA. USENIX Association.

Lee, T. B. (2013). Here's everything we know about PRISM to date. *The Washington Post*.

Libert, T. (2015a). Exposing the hidden web: Third-party http requests on one million websites. *International Journal of Communication*.

Libert, T. (2015b). Privacy implications of health information seeking on the web. *Communications of the ACM*.

Madden, M. (2014). Public perceptions of privacy and security in the post-snowden era. *Pew Research Center*.

Madden, M., Cortesi, S., Glasser, U., Lenhart, A., and Duggan, M. (2012). Parents, teens, and online privacy. *Pew Internet and American Life Project*.

Marthews, A. and Tucker, C. E. (2015). Government surveillance and internet search behavior. *SSRN*.

Mayer, J. R. and Mitchell, J. C. (2012). Third-party web tracking: Policy and technology. In *IEEE Symposium on Security and Privacy*, pages 413–427. IEEE.

Mayer, J. R. and Narayanan, A. (2011). Do not track: A universal third-party web tracking opt out. *Internet Engineering Task Force*, http://www.ietf.org/archive/id/draft-mayer-do-not-track-00.txt.

McDonald, A. M. and Cranor, L. F. (2008). The cost of reading privacy policies. *I/S: A Journal Of Law And Policy For The Information Society*, 4:543.

McDonald, A. M., Reeder, R. W., Kelley, P. G., and Cranor, L. F. (2009). A comparative study of online privacy policies and formats. In *Privacy Enhancing Technologies*, pages 37–55. Springer.

Milne, G. R., Culnan, M. J., and Greene, H. (2006). A longitudinal assessment of online privacy notice readability. *Journal of Public Policy & Marketing*, 25(2):238–249.

Mitchell, A. and Rosenstiel, T. (2015). State of the news media 2015. *Pew Research Center. Journalism & Media*.

National Institutes of Health, History of Medicine Division (2002). Greek medicine. *http://www.nlm.nih.gov/hmd/greek/greek_oath.html*.

National Security Agency (2013). Tor stinks presentation. *The Guardian*, http://www.theguardian.com/world/interactive/2013/oct/04/tor-stinks-nsa-presentation-document.

Network Advertising Initiative (2011). NAI Code of Conduct. *Network Advertising Initiative*.

Newman, N., Fletcher, R., Levy, D. A. L., and Nielsen, R. K. (2016). Reuters Institute Digital News Report 2016.

Nikiforakis, N., Kapravelos, A., Joosen, W., Kruegel, C., Piessens, F., and Vigna, G. (2013). Cookieless monster: Exploring the ecosystem of web-based device fingerprinting. In *IEEE Symposium on Security and Privacy*.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review*, 79(1).

Nissenbaum, H. F. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Penney, J. W. (2016). Chilling effects: Online surveillance and wikipedia use. *Berkeley Technology Law Journal*, 31:117.

PhantomJS (2016). Phantomjs is a headless webkit scriptable [browser] with a javascript api. it has fast and native support for various web standards: Dom handling, css selector, json, canvas, and svg. *http://phantomjs.org/*.

Pickard, V. (2014). *America's Battle for Media Democracy*. Cambridge University Press.

Reidenberg, J. R., Breaux, T., Cranor, L. F., French, B., Grannis, A., Graves, J. T., Liu, F., McDonald, A. M., Norton, T. B., Ramanath, R., et al. (2014). Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal*.

Rindfleisch, T. C. (1997). Privacy, information technology, and health care. *Communications of the ACM*, 40(8):92–100.

Roesner, F., Kohno, T., and Wetherall, D. (2012). Detecting and defending against third-party tracking on the web. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 12–12. USENIX Association.

Rotenberg, M. (2001). Fair information practices and the architecture of privacy (what Larry doesn't get). *Stanford Technology Law Review*, page 1.

Rothstein, M. A. and Talbott, M. K. (2006). Compelled disclosure of health information: Protecting against the greatest potential threat to privacy. *Journal of the American Medical Association*, 295(24):2882–2885.

Schneier, B. (2013). Surveillance as a business model. *Schneier on Security*.

Singel, R. (2008). Newly declassified files detail massive fbi data-mining project.

Solove, D. J. (2006). A taxonomy of privacy. *University of Pennsylvania Law Review*, pages 477–564.

Solove, D. J. (2012). Introduction: Privacy self-management and the consent dilemma. *Harvard Law Review*, 126:1880.

Soltani, A., Peterson, A., and Gellman, B. (2013). NSA uses Google cookies to pinpoint targets for hacking. *The Washington Post*, https://www.washingtonpost.com/blogs/the-switch/wp/2013/12/10/nsa-uses-google-cookies-to-pinpoint-targets-for-hacking.

Staff of Chairman Rockefeller (2013). A review of the data broker industry: Collection, use, and sale of consumer data for marketing purposes. *US Senate*.

Starr, P. (1999). Health and the right to privacy. *American Journal of Law & Medicine*, 25:193.

Stoycheff, E. (2016). Under surveillance examining Facebook's spiral of silence effects in the wake of NSA internet monitoring. *Journalism & Mass Communication Quarterly*, page 1077699016630255.

Supreme Court, U. S. (2011). United States v. Jones. *USA*.

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54.

Terry, N. P. (2003). Privacy and the health information domain: Properties, models and unintended results. *European Journal of Health Law*, 10(3):223–237.

Turow, J. (2003). *Americans & online privacy: The system is broken*. Annenberg Public Policy Center, University of Pennsylvania.

Turow, J. (2008). Niche envy: Marketing discrimination in the digital age. *MIT Press Books*.

Turow, J. (2012). *The daily you: How the new advertising industry is defining your identity and your worth*. Yale University Press.

Turow, J., Bleakley, A., Bracken, J., Carpini, M. X. D., Draper, N. A., Feldman, L., Good, N., Grossklags, J., Hennessy, M., Hoofnagle, C. J., et al. (2014). Americans, marketers, and the internet: 1999-2012. *Annenberg Public Policy Center*.

Turow, J. and Hennessy, M. (2007). Internet privacy and institutional trust insights from a national survey. *New Media & Society*, 9(2):300–318.

Turow, J., Hennessy, M., and Draper, N. A. (2015). The tradeoff fallacy, how marketers are misrepresenting American consumers and opening them up to exploitation. *The Annenberg School for Communication, University of Pennsylvania*.

Turow, J., King, J., Hoofnagle, C. J., Bleakley, A., and Hennessy, M. (2009). Americans reject tailored advertising and three activities that enable it. *University of Pennsylvania Scholarly Commons*.

Ur, B., Leon, P. G., Cranor, L. F., Shay, R., and Wang, Y. (2012). Smart, useful, scary, creepy: perceptions of online behavioral advertising. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, page 4. ACM.

US Federal Trade Commission (2010a). Complying with the FTC's Health Breach Notification Rule. *http://www.business.ftc.gov/documents/bus56-complying-ftcs-health-breach-notification-rule*.

US Federal Trade Commission (2010b). Protecting consumer privacy in an era of rapid change preliminary staff report. *http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-bureau-consumer-protection-preliminary-ftc-staff-report-protecting-consumer/101201privacyreport.pdf*.

Westin, A. (1967). *Privacy and Freedom*. Bodley Head, London.

Westin, A. F. (2003). Social and political dimensions of privacy. *Journal of Social Issues*, 59(2):431–453.

Yahoo! (2015). Microsoft and Yahoo Agree to Amend Search Partnership.

Yen, T.-F., Xie, Y., Yu, F., Yu, R. P., and Abadi, M. (2012). Host fingerprinting and tracking on the web: Privacy and security implications. In *Proceedings of the Network and Distributed System Security Symposium*.