INFORMS Journal on Computing

Now You See it, Now You Don't: Obfuscation of Online Third-Party Information Sharing

Ashkan Eshghi
Haskayne School of Business, University of Calgary, Calgary, Alberta T2N 1N4, Canada, ashkan.eshghi1@ucalgary.ca

Ram Gopal
Warwick Business School, University of Warwick, ram.gopal@wbs.ac.uk

Hooman Hidaji, Raymond Patterson
Haskayne School of Business, University of Calgary, Calgary, Alberta T2N 1N4, Canada,
hooman.hidaji@haskayne.ucalgary.ca, raymond.patterson@ucalgary.ca

The practice of sharing online user information with external third-parties has become the focal point of privacy concerns for consumer advocacy groups and policy makers. We explore the decisions by websites regarding the obfuscation that they employ to make it difficult for users to discover the extent of information sharing. Using a Bayesian model, we shed light on the websites' incentive to obfuscate user information sharing. We find that as content sensitivity increases, a website reduces its level of obfuscation. Further, more popular websites engage in higher levels of obfuscation than less popular ones. We provide an empirical analysis of obfuscation and user information sharing in News (low content sensitivity) and Health (high content sensitivity) websites and confirm key results from our analytical model. Our analysis illustrates that obfuscation of information sharing is a viable strategy that websites use to improve their profits.

Keywords: third-parties, information sharing, obfuscation, privacy, information asymmetry

> *"Unregulated free markets rarely reward the different kind of heroism, of those who restrain themselves from taking advantage of customers' psychological or informational weaknesses."*
>
> Akerlof and Shiller (2016)

1. Introduction

Websites are the crux of the Internet and online activity. A vast array of services and content provided by the websites are made possible through third-parties. Third-parties provide a wide range of services and resources to websites, ranging from basic functionality such as hosting content to monetization capabilities such as advertising. The website shares user information with third-parties when rendering the service, which enables it to monetize the users. This can include basic browsing data such as IP addresses and browser and device information, as well as personal information such as marital status, income, and interests. As users have become aware of the implications of the lack of online privacy, third-parties have become an important point of argument among the public, policy makers, and businesses. It is therefore crucial to understand the implications of the use of third-parties for privacy, and the mechanisms that affect the use of third-parties as well as the transparency of their use.

There is information asymmetry between users and websites in the sense that the website's user information sharing is not always transparent, and thus users cannot readily determine the level of information sharing with third-parties. This information asymmetry between website and users creates an opportunity for the website to take advantage of users' informational weaknesses and utilize strategies that favor their own interest (refer to Akerlof (1970), Hölmstrom (1979), Jenson and Meckling (1976), Arrow (1985), and Eisenhardt (1989)). While the extent of information sharing is websites' private information, users can expend some effort and monitor websites' sharing behavior to reduce or remove the information asymmetry. Depending on the benefits from information asymmetry, which are mediated by content sensitivity and user privacy concern, it may be profitable for the website to hide its information sharing behavior to mislead users and increase the cost of monitoring. We refer to this as *obfuscation*. According to Akerlof and Shiller (2015), uninformed online users, or "Information Phools" make decisions based on information that is intentionally crafted to mislead them.

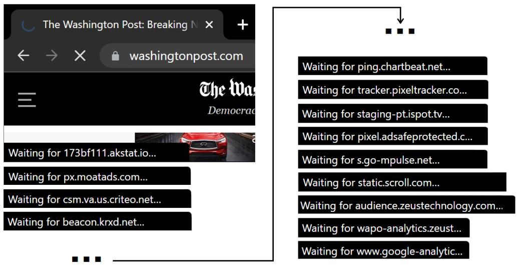
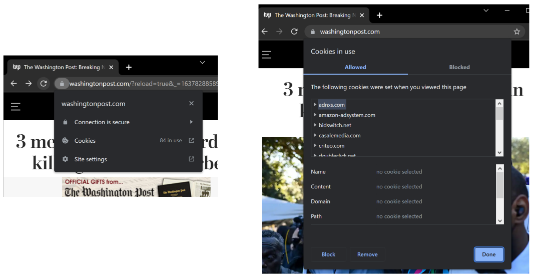
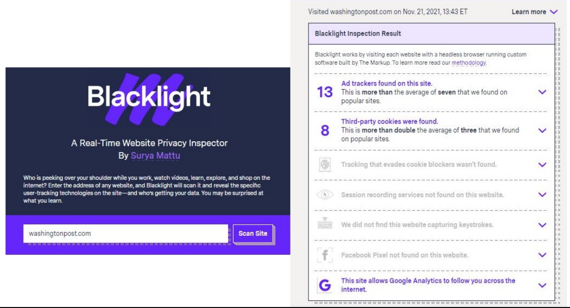
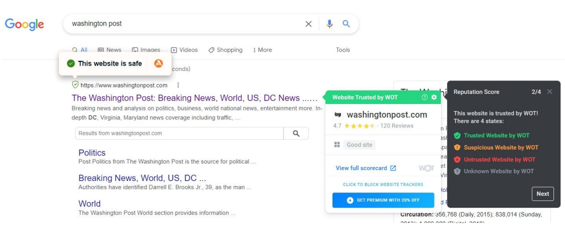
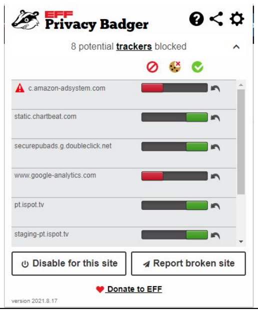
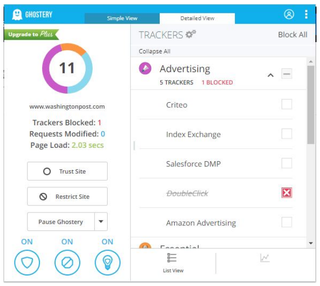
Focusing on websites' obfuscation of user information sharing with third-parties, we address the following research questions: 1) What are the website's incentives to obfuscate their sharing of user information with third-parties? 2) What type of websites are more prone to obfuscating their sharing of user information? To answer these questions, we first develop an analytical model based on the Bayesian equilibrium framework to capture and analyze the information asymmetry problem in the relationship between users and websites. We consider a monopolist website that is free to users, but charges third-parties for access to user information, and determines the level of obfuscation that it deploys. Our model sheds light on website's incentives to obfuscate user information sharing. Interestingly, we find that website's obfuscation decreases as content sensitivity increases. Additionally, obfuscation increases as the website's value to users increases, implying that prominent websites are more prone to obfuscation. Further, we verify the key findings of our model through an empirical analysis of data from 400 News and Health websites. Our empirical analysis, consistent with our model, shows that content sensitive websites (Health) do not obfuscate as much as less sensitive websites (News), and that prominent websites obfuscate more than non-prominent ones.

This paper contributes to our understanding of the websites' incentives, including those that drive websites to obfuscate their information sharing with third-parties, which has important implications for the information asymmetry between websites and users (Akerlof and Shiller 2015). Data regulation and privacy policy-makers can use our findings to design better policies which consider the websites' incentives to reduce the transparency of user information sharing. By clarifying the driving forces of website obfuscation, we contribute to the streams of literature on website monetization strategies (Gopal et al. 2018), online privacy (Li 2012, Pavlou 2011), obfuscation (Ellison and Wolitzky 2012, Gu and Wenzel 2014), and information flow and diffusion (Bai et al. 2012).

It is useful to provide some context for the information asymmetry problem between a website and its users, including privacy concerns, sharing of user information with third-parties, and its obfuscation by websites. Users have been shown to take action to alleviate their online privacy concerns. According to Cisco (2021), in 2021, 86% of online users care about their data privacy and 79% of those who care are willing to act, with 47% of these users actually acting to preserve their data privacy. Thus, 32% of users took active steps to maintain their online privacy, increasing by 3% from 2020 (Cisco 2021). Concerns over privacy are found to have caused 48% of respondents from "a survey of over 8,000 consumers and business buyers across 16 countries" to either stop buying from a company or using their service (Salesforce 2019). In the context of websites, users can carefully observe visible privacy signals (e.g., privacy policies and safety badges). However, a survey conducted in 2019, shows that only 8% of US adults understand privacy policies (PewResearch 2019). A research by Cisco (2021) shows that 36% of online users lack trust that companies are truthfully following their stated privacy policies.

Users can utilize monitoring tools to understand the level of information sharing and make decisions that protect their online privacy. There are privacy tools and services that give some indication of the abusive behavior of websites. These privacy tools either depend on a community of users to monitor and tag abusive behavior (e.g., Web of Trust, Webutation, and Avast Web Reputation Plugin) or keep track of the third-parties that are engaged by the website (e.g., Blacklight, Lightbeam, Privacy Badger, and Ghostery). As an indication of widespread use of privacy tools, one of these tools, Web of Trust, has more than 140 million users (TechTimes 2020). Some examples of privacy tools and services are presented in Table 1.

Table 1 Examples of Using Privacy Tools for an Example Website (`washingtonpost.com`)

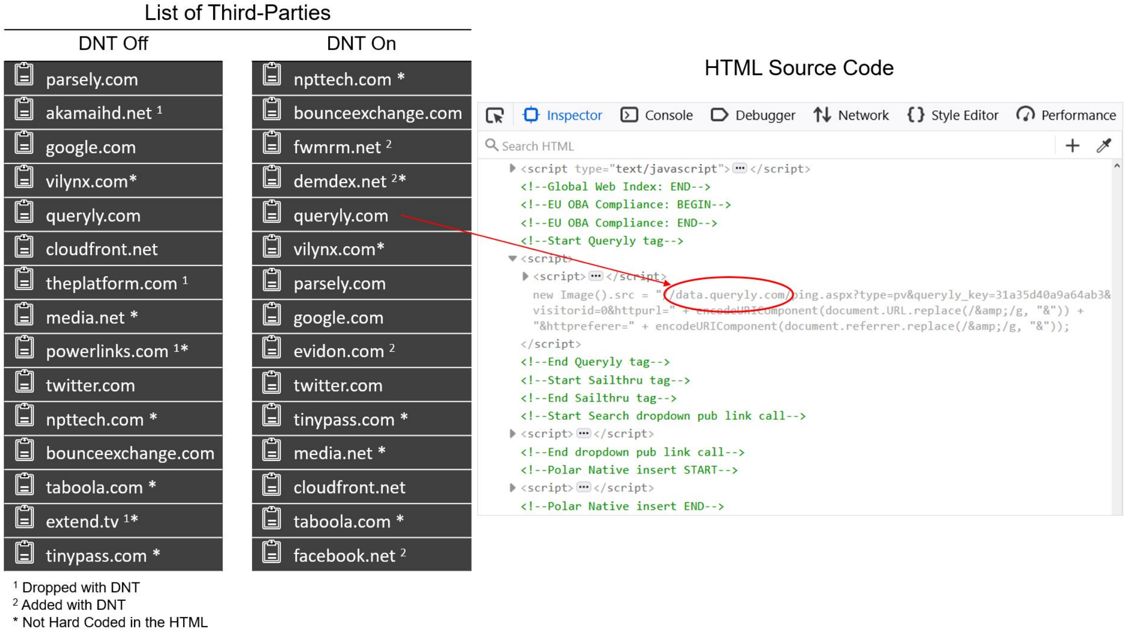
Browser loading notifications	Google Chrome site information
 The Washington Post: Breaking N washingtonpost.com Waiting for 173bf111.akstat.io... Waiting for px.moatads.com... Waiting for csm.va.us.criteo.net... Waiting for beacon.krxd.net... Waiting for ping.chartbeat.net... Waiting for tracker.pixeltracker.co... Waiting for staging-pt.ispot.tv... Waiting for pixel.adsafeprotected.c... Waiting for s.go-mpulse.net... Waiting for static.scroll.com... Waiting for audience.zeustechnology.com... Waiting for wapo-analytics.zeust... Waiting for www.google-analytic...	 washingtonpost.com/?reload=true&_=16378288589 washingtonpost.com Connection is secure Cookies 84 in use Site settings Cookies in use Allowed Blocked The following cookies were set when you viewed this page adnxs.com amazon-adsystem.com bidswitch.net casalemedia.com criteo.com doubleclick.net Name no cookie selected Content no cookie selected Domain no cookie selected Path no cookie selected Block Remove Done
Blacklight	**Web of Trust – Avast**
 Blacklight A Real-Time Website Privacy Inspector By Surya Mattu Who is peeking over your shoulder while you work, watch videos, learn, explore, and shop on the internet? Enter the address of any website, and Blacklight will scan it and reveal the specific user-tracking technologies on the site—and who's getting your data. You may be surprised at what you learn. washingtonpost.com Scan Site Visited washingtonpost.com on Nov. 21, 2021, 13:43 ET Learn more Blacklight Inspection Result Blacklight works by visiting each website with a headless browser running custom software built by The Markup. To learn more read our methodology. 13 Ad trackers found on this site. This is **more than** the average of **seven** that we found on popular sites. 8 Third-party cookies were found. This is **more than double** the average of **three** that we found on popular sites. Tracking that evades cookie blockers wasn't found. Session recording services not found on this website. We did not find this website capturing keystrokes. Facebook Pixel not found on this website. This site allows Google Analytics to follow you across the internet.	 washington post All News Images Videos Shopping More Tools This website is safe https://www.washingtonpost.com The Washington Post: Breaking News, World, US, DC News ...... Breaking news and analysis on politics, business, world national news, entertainment more. In-depth **DC**, Virginia, Maryland news coverage including traffic, ... Results from washingtonpost.com Politics Post Politics from The Washington Post is the source for political ... Breaking News, World, US, DC ... Authorities have identified Darrell E. Brooks Jr., 39, as the man ... World The Washington Post World section provides information ... Website Trusted by WOT washingtonpost.com 4.7 ★★★★★ · 120 Reviews Good site View full scorecard CLICK TO BLOCK WEBSITE TRACKERS GET PREMIUM WITH 20% OFF Reputation Score 2/4 This website is trusted by WOT! There are 4 states: Trusted Website by WOT Suspicious Website by WOT Untrusted Website by WOT Unknown Website by WOT Next Circulation: 356,768 (Daily, 2015); 838,014 (Sunday,
Privacy Badger	**Ghostery**
 EFF Privacy Badger 8 potential trackers blocked c.amazon-adsystem.com static.chartbeat.com securepubads.g.doubleclick.net www.google-analytics.com pt.ispot.tv staging-pt.ispot.tv Disable for this site Report broken site Donate to EFF version 2021.8.17	 GHOSTERY Simple View Detailed View Upgrade to Plus 11 www.washingtonpost.com Trackers Blocked: 1 Requests Modified: 0 Page Load: 2.03 secs Trust Site Restrict Site Pause Ghostery ON ON ON TRACKERS Block All Collapse All Advertising 5 TRACKERS 1 BLOCKED Criteo □ Index Exchange □ Salesforce DMP □ ~~DoubleClick~~ ☒ Amazon Advertising □ List View

On the other hand, websites can use obfuscation techniques to make monitoring more difficult. Websites use a variety of obfuscation techniques to impede users from understanding the true level of information sharing and its privacy implications. Obfuscation refers to practices and strategies that firms employ to confuse users and prevent them from recognizing the best offer (Ellison and Wolitzky 2012 and Gu and Wenzel 2014). One common technique employed by websites is use of "dark patterns", which tricks users into giving away their privacy (Vincent 2021). Such efforts by websites make it hard for users to discover the true level of information sharing and are instantiations of obfuscation strategies. In our context, obfuscation involves strategically preventing users from recognizing the true level of information sharing with third-parties.

For an example of website obfuscation of third-parties, consider the following. Users can utilize passive privacy tools, which signal to the website that such users are concerned about their privacy. One popular such tool is the passive request not to be tracked, known as a Do Not Track (DNT) request. Users may perceive the website's level of information sharing by comparing the number of third-parties used when the website is visited with and without a DNT request. On the other hand, websites can obfuscate the sharing of user information by making it hard for users to discern the reaction of websites to explicit user privacy requests. For example, some websites both add and drop third-parties that they share data with when users ask not to be tracked. This creates confusion about the true intent of the website and thus obfuscates the website's information sharing behavior. Moreover, websites can dynamically change the third-parties that are hard coded in the HTML source code. This creates even more confusion and makes it harder for privacy services to determine the true level of information sharing. To illustrate such obfuscation practices, we present the response of a website to a DNT request in Figure 1.

As shown in Figure 1, for this particular website, 15 third-parties are utilized. When the user enables DNT, the website drops four third-parties and replaces them with four new third-parties, leaving eleven third-parties unchanged. Adding to this state of confusion is the question of whether the third-parties can be discovered through the HTML source code of the website. In this illustration we find that 7 third-parties are not hard coded in the HTML source code when DNT is off, and 6 are not included in the HTML source code when DNT is on. This implies that if a user or an organization were to study the HTML for this website, they would not be able to discover all third-parties that are utilized. All

Figure 1 An Example of Website Information Sharing Obfuscation



List of Third-Parties

DNT Off	DNT On
parsely.com	npttech.com *
akamaihd.net [1]	bounceexchange.com
google.com	fwmrm.net [2]
vilynx.com*	demdex.net [2]*
queryly.com	queryly.com
cloudfront.net	vilynx.com*
theplatform.com [1]	parsely.com
media.net *	google.com
powerlinks.com [1]*	evidon.com [2]
twitter.com	twitter.com
npttech.com *	tinypass.com *
bounceexchange.com	media.net *
taboola.com *	cloudfront.net
extend.tv [1]*	taboola.com *
tinypass.com *	facebook.net [2]

[1] Dropped with DNT
[2] Added with DNT
* Not Hard Coded in the HTML

HTML Source Code

```
<script type="text/javascript">…</script>
<!--Global Web Index: END-->
<!--EU OBA Compliance: BEGIN-->
<!--EU OBA Compliance: END-->
<!--Start Queryly tag-->
<script>
  <script>…</script>
  new Image().src = "//data.queryly.com/ping.aspx?type=pv&queryly_key=31a35d40a9a64ab3&
  visitorid=0&httpurl=" + encodeURIComponent(document.URL.replace(/&amp;/g, "&")) +
  "&httpreferer=" + encodeURIComponent(document.referrer.replace(/&amp;/g, "&"));
</script>
<!--End Queryly tag-->
<!--Start Sailthru tag-->
<!--End Sailthru tag-->
<!--Start Search dropdown pub link call-->
<script>…</script>
<!--End dropdown pub link call-->
<!--Polar Native insert START-->
<script>…</script>
<!--Polar Native insert END-->
```

of this represents a great deal of change in reaction to the DNT request, which confuses users and privacy services about the true intention of the website.

2. Literature Review

The technical implications of third-parties and their impact on users' information diffusion and leakage have been studied using a variety of methods such as crowd-sourcing (Yu et al. 2016) and web crawling (Englehardt and Narayanan 2016). Krishnamurthy et al. (2011) study the websites with user registration, and find 75% of the websites to leak sensitive user information to third-parties. Roesner et al. (2012) detect and classify five different types of third-party trackers based on how they work within the browser environment. Acar et al. (2020) study the extent of data collection by third-parties, investigating the scripts that are directly embedded on web pages. We extend this literature by studying the website's obfuscation of third-party usage among websites. Our analysis shows that obfuscation is popular among popular websites.

Users' browsing and personal data is shared with third-parties by websites, mainly for the purpose of monetization and advertising. This is done through use of third-party cookies and passing data through *https* requests (Libert 2015, Englehardt et al. 2015, Englehardt and Narayanan 2016). There have been several studies on the relationship between websites and third-parties, and the underlying mechanisms for websites to improve their usability (Gupta et al. 2007) and monetization (Gopal et al. 2018) through third-party trackers. Gopal et al. (2018) study the website's trade-off between sharing user data with third-parties and user privacy concerns. Gupta et al. (2007) propose a methodology to improve the linkage of websites according to user preferences estimated based on user data, and Yang et al. (2013) offer a framework to discover users online shopping patterns across websites using their online behavior. The interaction of websites and third-parties forms an integral part of the website's monetization of users, which we study in this paper. Rather than focusing on the operational details of websites in terms of their linkage and discovery of users, we focus on the incentive of websites to obfuscate their use of third-parties.

Use of third-parties comes with implications for privacy. Online privacy has been a subject of many prior studies. Smith et al. (2011), Pavlou (2011), and Li (2012) provide a comprehensive review of the extensive online information privacy literature and develop frameworks for theoretical research on user privacy decision making. Bai et al. (2012) consider the security and privacy issues at organizational workflows and suggest consideration to improve exposure. We directly include privacy in our analysis, and extend it to include the interaction between privacy concern of users, content sensitivity of websites, and the level of information sharing at the website. This allows us to capture the different factors that drive the total privacy cost that users face.

Information asymmetry surrounds our analysis of obfuscation. Given that users do not readily know the extent of information sharing, there is information asymmetry between website and users. Websites can utilize this to their benefit. Such implications of information asymmetry have been extensively studied, for example in the context of quality and uncertainty (Akerlof 1970), moral hazard (Hölmstrom 1979), and agency (Jenson and Meckling 1976, Arrow 1985, Eisenhardt 1989). We extend this by including the ability of websites to change the level of information asymmetry, which is done through obfuscation.

Obfuscation encompasses strategies to increase the information asymmetry between the website and users by making it costly to discover the true extent of website information sharing with third-parties. Researchers have studied obfuscation as a way for firms to prevent customers from recognizing the best offer (Ellison and Wolitzky 2012, Gu and Wenzel 2014), and this is consistent with our view of obfuscation by websites. Obfuscation of information has also been studied in the context of firms' information sharing with stakeholders within firm disclosures (Bushee et al. 2018), higher education signaling (Goldstein and Eaton 2021), and product and price search results (Wilson 2010). However, to the best of our knowledge, our analysis is the first to consider the obfuscation carried out by websites to make their sharing of user information less transparent.

3. Analytical Model

In this section, we develop an analytical model to study the website-user relationship and analyze the website's decision regarding obfuscation of information sharing with third-parties. For convenience, a summary of our analytical model notations are shown in Table 2.

We consider a website that offers a free service to potential users, and charges third-parties for access to user information. We assume that the website is one of two types with respect to its level of information sharing, denoted as $t \in \{L, H\}$: low-type (L), or high-type (H). We assume there is a common belief that the website is low-type (L) with probability

Table 2 Variables

Notation	Definition
X	Value of the service that is provided by the website.
r	Individual IT illiteracy, which is defined as user's inability to discover website's true information sharing level, where $r \in [0,1]$.
s	Privacy sensitivity of website content (i.e., content sensitivity).
t	Website type with respect to its level of information sharing, where $t \in \{L, H\}$, $t = L$ is low, and $t = H$ is high.
v	Individual user's privacy concern, implying their concern for their information being shared with third-parties, where $v \in [\underline{v}, \overline{v}]$, $0 \leqslant \underline{v} < \overline{v}$.
m_t	Level of information sharing for website of type t, where $m_L < m_H$.
η_t	Obfuscation level of website of type t.
θ	Probability that website's type t is L, whereas probability that website's type t is H is $1-\theta$.
$K_{v,r}$	Action of a user with privacy concern v and IT illiteracy r, where $K_{v,r} \in \{u, e, n\}$ and u denotes using the website without expending effort to discover the website, e denotes expending effort, and n denotes neither using the website nor expending effort.
$U^K(v, r, t)$	The utility from action K that a user with privacy concern v and IT illiteracy r gains from website type t.
$C(\eta_t)$	Obfuscation cost function.
$D_t(\eta_L, \eta_H)$	Demand that type t website obtains by choosing η_t.
Π_t	Profit of type t website.

θ and high-type (H) with probability $1-\theta$. The website's level of information sharing is the website's private information, but users can expend effort to determine this.

We define obfuscation as the practice of the website to make it difficult for users to determine its type. If the website does not obfuscate and is fully transparent, the cost of effort to discover the website type (its level of information sharing) is zero. As the website increases obfuscation, it increases the cost of user discovery of the website type. The cost to the user of discovering the website type is a function of both the user's technical ability and the level of obfuscation employed by the website. The more ability a user has, the lower their cost of discovery, and vice-versa. We refer to a users' lack of technical ability as their *IT illiteracy*, denoted as r, which is assumed to be uniformly distributed $(r \in [0,1])$ across the population. The effort cost needed to discover the website type for a user with IT illiteracy r for website of type t is given as $r\eta_t$, where η_t is the obfuscation level of website of type t. Where $r = 0$, the cost of discovery is zero, corresponding to the lowest user IT illiteracy. Where $r = 1$, the effort needed to discover the website type is η_t, corresponding to the highest level of user IT illiteracy.

Utility that users gain from using the website depends on the value of the service, denoted as X, and the privacy cost of using the service. The privacy cost of using the website for users depends on the website's level of information sharing (website type, t), website content sensitivity (denoted as s), and users' privacy concern (denoted as v). The website's level of information sharing is denoted as m_t for a website of type t, where $m_L < m_H$. Therefore, the total privacy cost that a user with privacy concern v incurs if they use the website of type t with content sensitivity of s is given as svm_t.

Content sensitivity (s) relates to the website's content. It is believed that different types of websites have different levels of content sensitivity, where, for example, health-related information is more sensitive than news (Gopal et al. 2018). User privacy concern (v), on the other hand, is specific to a given user, and implies their concern for their information being shared with third-parties. Those with high privacy concern are more sensitive to their data being shared (Gopal et al. 2018). Users are heterogeneous with respect to their privacy concern (Chellappa and Shivendu 2007), and are assumed to have a privacy concern according to a uniform distribution $v \in [\underline{v}, \overline{v}]$. Note that our model accounts for two dimensions of user heterogeneity: IT illiteracy (r) and privacy concern (v). We assume these two dimensions to be independent of each other among users.[1]

We model the website-user relationship as an incomplete information game. The website chooses the level of obfuscation and simultaneously users decide to either use the website,

[1] This assumption does not qualitatively impact our result, as any correlation between user privacy concern and IT illiteracy can be superimposed on our findings to capture the results for any special case.

expend effort to discover the website type or not use the website. As discussed above, users choose from three possible choices:

1. Use the website without expending effort. We denote these users with u.

2. Expend effort to discover the website type (t). Given the website type, these users use the website if they receive positive utility from the website. We denote these users with e.

3. Do not use the website and do not expend effort. We denote these users with n.

Accordingly, the users' action set is $\{u, e, n\}$, and each individual user's action is denoted as $K_{v,r} \in \{u, e, n\}$. The utility that users gain from choosing each action for each website type is given as:

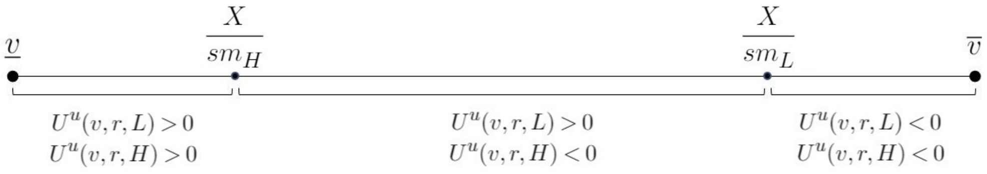
$$\begin{aligned} &U^u(v,r,t) = X - svm_t \\ &U^e(v,r,t) = \begin{cases} X - svm_t - r\eta_t & \forall\, v \leqslant \frac{X}{sm_t} \\ -r\eta_t & \forall\, v > \frac{X}{sm_t} \end{cases} \\ &U^n(v,r,t) = 0 \end{aligned} \tag{1}$$

Users with privacy concern $v \leqslant X/sm_H$ gain positive utility from using the website irrespective of the website's level of obfuscation and their IT illiteracy ($U^u(v,r,L) > 0$, $U^u(v,r,H) > 0$). Therefore, these users use the website without expending effort. On the other hand, users with $v \geqslant X/sm_L$ gain negative utility from using the website irrespective of the website's level of obfuscation and their IT illiteracy ($U^u(v,r,L) < 0$, $U^u(v,r,H) < 0$), and therefore, do not use the website and do not expend effort.[2] The analysis for users with $X/sm_H < v < X/sm_L$ (where $U^u(v,r,L) > 0$, $U^u(v,r,H) < 0$) is more complicated, as these users need to decide whether to use the website, not use the website, or expend

[2] For brevity, we focus on the more interesting scenario where $\underline{v} < X/sm_H < X/sm_L < \overline{v}$. Our results also apply to the case where this assumption does not hold.

effort to identify the website type, which then determines whether they use or not use the website. These decisions depend on the expected utility that these users receive from the website, as we explain next. Figure 2 provides an overview of the users' utility for users with different privacy concerns.

Figure 2 Users' Utility from Using the Website

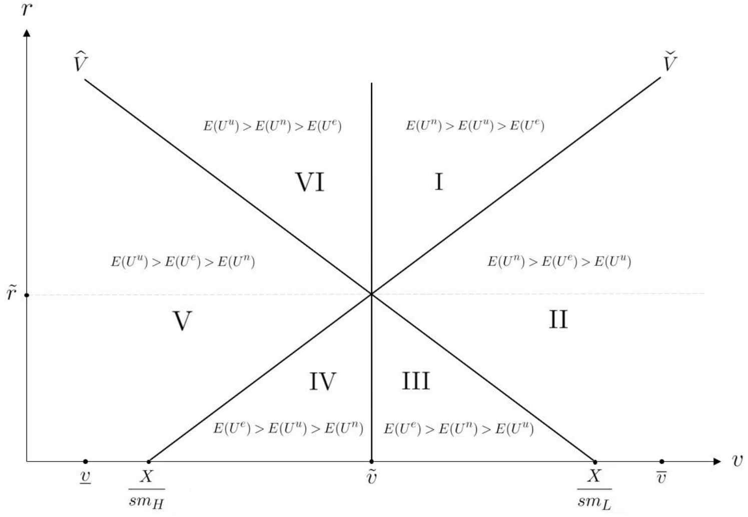


Based on (1), the expected utility of users (over the two possible website types) with $X/sm_H < v < X/sm_L$ is derived as a function of the expected level of information sharing, $E(m) = \theta m_L + [1-\theta]m_H$, and the expected level of obfuscation, $E(\eta) = \theta\eta_L + [1-\theta]\eta_H$. Using these definitions, we can derive the expected user utility as:

$$\begin{aligned} &E(U^u(v,r)) = \theta U^u(v,r,L) + [1-\theta]U^u(v,r,H) = X - svE(m) \\ &E(U^e(v,r)) = \theta U^e(v,r,L) + [1-\theta]U^e(v,r,H) = \theta[X - svm_L] - rE(\eta) \\ &E(U^n(v,r)) = 0 \end{aligned} \tag{2}$$

The expected user utility of use is simply the expected utility for the two website types. The expected utility of effort is composed of the weighted utility of using the website if it is low-type ($\theta[X - svm_L]$) subtracted by the expected cost of effort, noting that the utility of not using the website if it is high-type is zero. Based on the above utility expectations, we can characterize the users based on their actions. We denote the privacy concern of the users who are indifferent between using or not using, that is users with $E(U^u(v,r)) = E(U^n(v,r))$, as $\tilde{v} = X/sE(m)$. It can be seen that these indifferent users' privacy concern does not depend on their IT illiteracy r. We denote the privacy concern of users who are indifferent between using the website or expending effort to discover the website type, that is users with $E(U^u(v,r)) = E(U^e(v,r))$, as $\check{V}(r) = [[1-\theta]X + rE(\eta)]/[1-\theta]sm_H$. Finally, we denote the privacy concern of users who are indifferent between expending effort to discover the website type or not using the website, that is users with $E(U^e(v,r)) = E(U^n(v,r))$ as $\widehat{V}(r) = [\theta X - rE(\eta)]/\theta sm_L$. Note that the privacy concern of the users who are indifferent between using or expending effort ($\check{V}$) and not using or expending effort ($\widehat{V}$) characterized above, depends on their IT illiteracy (r). Therefore, $\check{V}$ and $\widehat{V}$ are functions of r and can be shown as indifference lines. Where the lines $v = \tilde{v}$, $\check{V}$, and $\widehat{V}$ intersect, we have $r = \tilde{r} = \theta[1-\theta]X(m_H - m_L)]/E(m)E(\eta)$. The indifference lines and their intersection points are illustrated in Figure 3.

Figure 3 Users' Expected Utility


r
$\widehat{V}$
$\check{V}$
$E(U^u) > E(U^n) > E(U^e)$
$E(U^n) > E(U^u) > E(U^e)$
VI
I
$E(U^u) > E(U^e) > E(U^n)$
$E(U^n) > E(U^e) > E(U^u)$
$\tilde{r}$
V
II
IV
III
$E(U^e) > E(U^u) > E(U^n)$
$E(U^e) > E(U^n) > E(U^u)$
v
$\underline{v}$
$\frac{X}{sm_H}$
$\tilde{v}$
$\frac{X}{sm_L}$
$\overline{v}$

Users' optimal decision can be represented as their best response to the website's level of obfuscation given the users' belief about the website type. Each user chooses the action $K_{v,r} \in \{u, e, n\}$ which maximizes her expected utility. For example, a given user uses the website if her expected utility of using the website is larger than the expected utility of both not using the website and expending effort to discover the website type. Accordingly, the users' best response function is defined as $B_{v,r}(\eta_L, \eta_H) = \underset{K_{v,r}}{\operatorname{argmax}} E(U^{K_{v,r}}(v,r))$. Comparing the expected utilities in (2), we can write the users' best response function as:

$$B_{v,r}(\eta_L, \eta_H) = \begin{cases} u, & \text{if} \quad v \leqslant \tilde{v} \ \& \ v \leqslant \check{V}(r) \\ e, & \text{if} \quad \check{V}(r) < v < \widehat{V}(r) \\ n, & \text{if} \quad v > \tilde{v} \ \& \ v \geqslant \widehat{V}(r) \end{cases} \tag{3}$$

As shown in Figure 3, in regions I and II the expected utility of not using the website, $E(U^n(v,r))$, is higher than both the expected utility of using the website, $E(U^u(v,r))$, and the expected utility of expending effort to discover the website type, $E(U^e(v,r))$. Therefore, the optimal decision for the users in these regions is to not use the website corresponding to $B_{v,r}(\eta_L, \eta_H) = n$. In regions III and IV the expected utility of expending effort to discover the website type is higher than both the expected utility of using and the expected utility of not using the website. Therefore, the optimal decision for the users in these regions is to expend effort to discover the website type, corresponding to $B_{v,r}(\eta_L, \eta_H) = e$. In regions V and VI the expected utility of using the website is higher than the expected utility of not using the website and the expected utility of expending effort to discover the website type. Therefore, the optimal decision for the users in these regions is to use the website, corresponding to $B_{v,r}(\eta_L, \eta_H) = u$.

3.1. Demand Specification

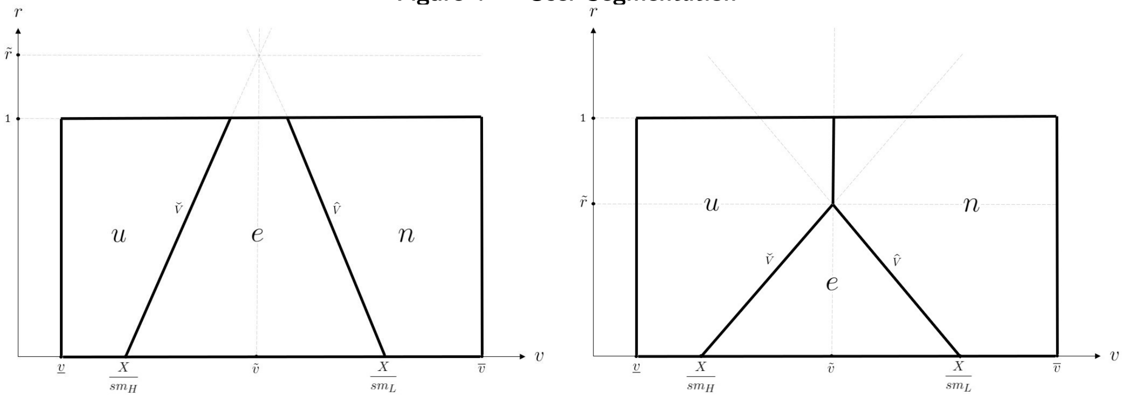
As obfuscation $(E(\eta))^3$ increases, the indifference lines between using and expending effort (the slope of $\check{V}$ with respect to r is $E(\eta)/[1-\theta]m_H$) and between expending effort and not

[3] Hereafter, we refer to expected obfuscation (composed of two possible website types' level of obfuscation) simply as obfuscation, as these two concepts are closely related. As we show in our analysis, a low-type website sets obfuscation to zero, and therefore, expected obfuscation in equilibrium is perfectly correlated with a high-type website's level of obfuscation.

using (the slope of $\widehat{V}$ with respect to r is $-E(\eta)/\theta m_L$) become more steep. Therefore, the crossing point of these two lines ($\tilde{r}$) moves down. We can specify demand depending on the obfuscation as compared to the threshold $\tilde{\eta} = \theta[1-\theta]X[m_H - m_L]/E(m)$. We discuss the significance of $\tilde{\eta}$ later in our analysis.

Figure 4 shows the user segments for the different expected levels of obfuscation. Where obfuscation is low ($E(\eta) < \tilde{\eta}$, Figure 4.(a)), the crossing point between $\check{V}$ and $\widehat{V}$ occurs above 1 ($\tilde{r} > 1$). In this case, there exist some users at any level of IT illiteracy r that have the incentive to expend effort to discover website type. As obfuscation increases to $E(\eta) = \tilde{\eta}$, the crossing point occurs at 1 ($\tilde{r} = 1$). In this case, the most IT illiterate user ($r = 1$) with an average privacy concern ($v = \tilde{v}$) is indifferent between using the website, not using the website, and expending effort to discover the website type. The threshold $\tilde{\eta}$ is the obfuscation level for which this special case occurs. As obfuscation increases even more ($E(\eta) > \tilde{\eta}$), the crossing point occurs at below 1 ($\tilde{r} < 1$). In this case, users with $\tilde{r} \leqslant r \leqslant 1$ do not have an incentive to expend effort to discover website type, and either use or not use the website irrespective of their privacy concern v.

Figure 4 User Segmentation



(a) $E(\eta) < \tilde{\eta}$ **(b)** $E(\eta) \geq \tilde{\eta}$

As explained above, only those users expend effort who know that given the cost of obfuscation, they gain positive utility from the website if they discover it to be low-type.

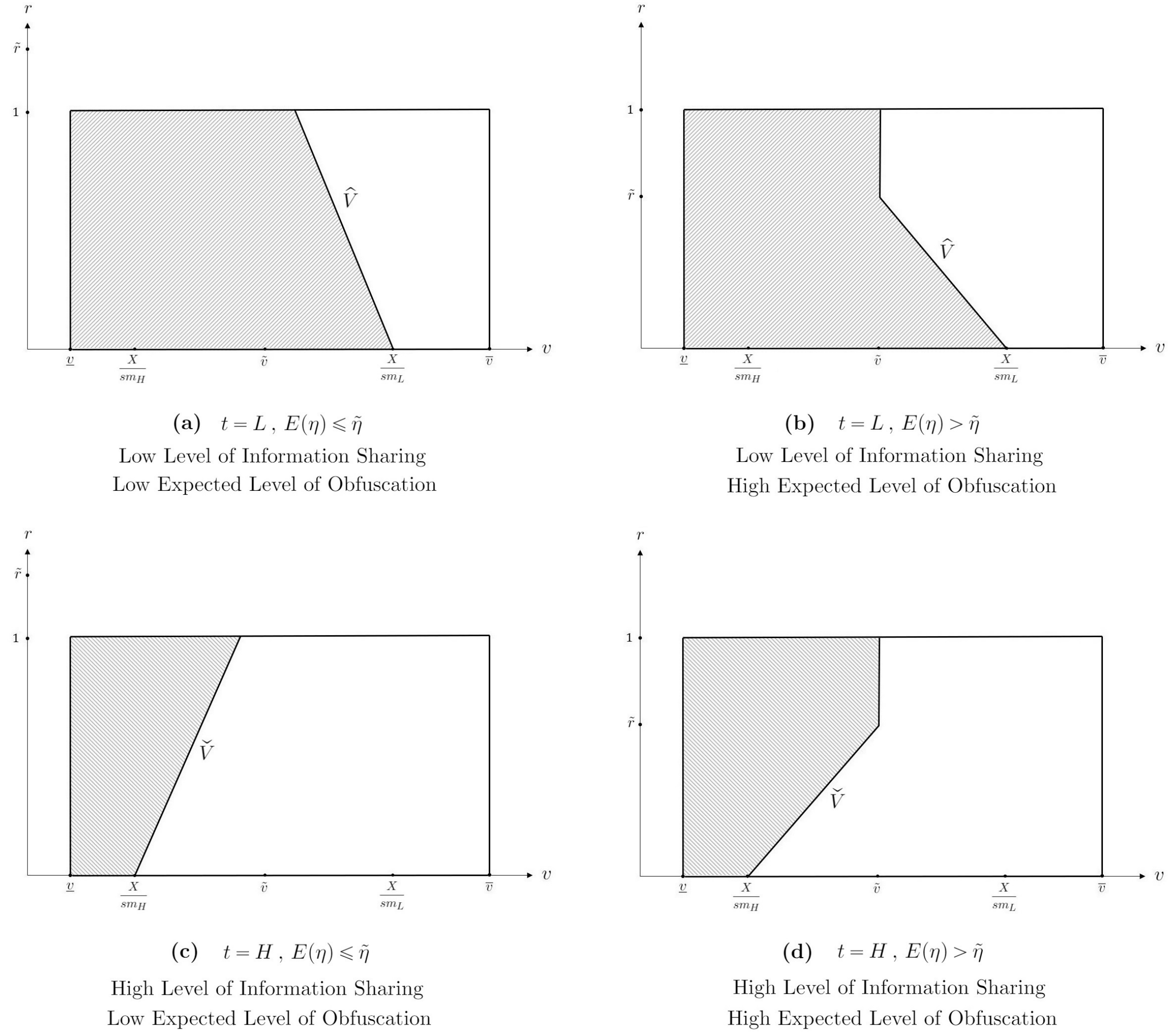
Therefore, users who expend effort, use the website only if the website is low-type, which implies that a low-type website's demand is composed of users who use the website without expending effort, and users who expend effort. The high-type website's demand, on the other hand, is composed only of users who use the website without expending effort, because if users expend effort and discover the true level of information sharing to be high, then they do not use the website. Given the users' best response function and the expected level of obfuscation, demand for each website type is characterized as:

$$\begin{aligned} D_L(\eta_L,\eta_H) &= \frac{\int_0^{\min\{\tilde{r},1\}} \widehat{V}\,dr + \int_{\min\{\tilde{r},1\}}^1 \tilde{v}\,dr - \underline{v}}{\overline{v}-\underline{v}} \\ D_H(\eta_L,\eta_H) &= \frac{\int_0^{\min\{\tilde{r},1\}} \check{V}\,dr + \int_{\min\{\tilde{r},1\}}^1 \tilde{v}\,dr - \underline{v}}{\overline{v}-\underline{v}} \end{aligned} \tag{4}$$

Considering the demand, the effect of obfuscation is that it makes it harder for users to discover the type of website, that is, it reduces users who expend effort. This shifts some of the users to not use the website, and others to use the website, both without expending effort. This increases the ratio of users who use the website, which improves the demand for the high-type website. On the other hand, this effect reduces the total ratio of users who either use the website or expend effort, which constitutes the low-type website's demand. The demand for each type of website is illustrated in Figure 5.

Because obfuscation impacts demand only through users who expend effort, users with $r > \tilde{r}$ are not impacted by obfuscation. This diminishes the impact of obfuscation on demand beyond $\tilde{\eta}$. Therefore, the impact of obfuscation on demand has two distinct pieces. Where obfuscation is low ($E(\eta) < \tilde{\eta}$, Figures 5.(a) and 5.(c)), obfuscation impacts the demand for users with any level of IT illiteracy ($r \leqslant 1$). However, where obfuscation is high ($E(\eta) > \tilde{\eta}$, Figures 5.(b) and 5.(d)), obfuscation impacts the demand only for IT literate users (users with $r < \tilde{r}$) who expend effort to discover website type. Therefore, in the second

Figure 5 Demand for Each Type of Website



(a) $t = L$, $E(\eta) \leqslant \tilde{\eta}$
Low Level of Information Sharing
Low Expected Level of Obfuscation

(b) $t = L$, $E(\eta) > \tilde{\eta}$
Low Level of Information Sharing
High Expected Level of Obfuscation

(c) $t = H$, $E(\eta) \leqslant \tilde{\eta}$
High Level of Information Sharing
Low Expected Level of Obfuscation

(d) $t = H$, $E(\eta) > \tilde{\eta}$
High Level of Information Sharing
High Expected Level of Obfuscation

piece where obfuscation is high, its impact on demand is diminished. This phenomenon has an important implication for the optimal level of obfuscation, as it makes the demand function piece-wise with respect to obfuscation (two pieces: $E(\eta) < \tilde{\eta}$ and $E(\eta) > \tilde{\eta}$). As we explain in the next section (Section 3.2), in the case where an interior solution exists, this drives the optimal level of obfuscation: in this case a high-type website increases the obfuscation beyond $\tilde{\eta}$, after which the marginal return of obfuscation is dampened, and the optimal level of obfuscation is reached.

3.2. Bayesian Equilibrium Analysis

As discussed, the website provides a free service to users, but charges third parties a fixed and exogenous price per user, which we normalize to 1. Therefore, website's revenue is given as $D_t(\eta_L, \eta_H)m_t$. The obfuscation cost function is $C(\eta_t)$, where $C'(\eta_t) > 0$, $C''(\eta_t) \geqslant 0$ and $C(0) = 0$. These standard assumptions imply that obfuscation is costly and that the website chooses easy obfuscation practices first, which makes obfuscation increasingly difficult. In other words, as website employs additional obfuscation practices, it becomes increasingly difficult to find new ways to prevent users from discovering the website type. We also allow for a linear cost function (where $C''(\eta_t) = 0$) to extend our results to this special case.

The website's profit function is given as:

$$\Pi_t = D_t(\eta_L, \eta_H)m_t - C(\eta_t) \tag{5}$$

Note that the website revenue is composed of the demand (D_t) times the level of information sharing (m_t, which equates to the per-user revenue). Content sensitivity and user privacy concern impact the website profit only indirectly through the demand. On the other hand, the level of information sharing (m_t) impacts revenue both directly as the per-unit revenue, and indirectly through demand (D_t), as users may refrain from using a website that extensively shares their data with third-parties.

The optimal decision of a type t website can be represented as its best response to users decisions. The website maximizes its profit function by choosing the level of obfuscation η_t. The best response function of a type t website in reaction to the collective set of all individual user actions $\{K_{v,r} | \underline{v} \leqslant v \leqslant \overline{v}\ ,\ 0 \leqslant r \leqslant 1\}$ is therefore given as:

$$B_t(\{K_{v,r} | \underline{v} \leqslant v \leqslant \overline{v}\ ,\ 0 \leqslant r \leqslant 1\}) = \underset{\eta_t}{\operatorname{argmax}}\, \Pi_t \tag{6}$$

The Bayesian Nash equilibria (Harsanyi 1968) of this game are vectors $(\eta_t^*, K_{v,r}^*)$ such that:

$$\begin{aligned} B_{v,r}(\eta_L^*, \eta_H^*) = K_{v,r}^* \quad &, \qquad \forall \ (v \in [\underline{v}, \overline{v}]\ , \ r \in [0,1]) \\ B_t(\{K_{v,r}^* |\ \underline{v} \leqslant v \leqslant \overline{v}\ ,\ 0 \leqslant r \leqslant 1\}) = \eta_t^* \quad &, \qquad \forall \ t \in \{L, H\} \end{aligned} \tag{7}$$

Considering the impact of obfuscation on website profit in (5), it can be seen that for a low-type website, there is no incentive to increase obfuscation, as it both decreases demand and increases the costs. However, if the positive effect of obfuscation on demand outsizes the negative effect of increased costs, a high-type website has an incentive to obfuscate. Our first lemma characterizes the equilibrium obfuscation level. The details for deriving the equilibrium and all proofs are relegated to the Appendix.

Lemma 1. **Level of Obfuscation in Equilibrium**

(a) A low-type website does not obfuscate ($\eta_L^* = 0$).

(b) Where the marginal cost of obfuscation at the point $\tilde{\eta}/[1-\theta]$ is low ($C'(\tilde{\eta}/[1-\theta]) < 1/2s[\overline{v} - \underline{v}]$), there is an interior solution for the high-type website's optimal level of obfuscation, where $\eta_H^* > \tilde{\eta}/[1-\theta]$.

(c) Where the marginal cost of obfuscation at the point $\tilde{\eta}/[1-\theta]$ is high ($C'(\tilde{\eta}/[1-\theta]) \geqslant 1/2s[\overline{v} - \underline{v}]$), the high-type website's optimal level of obfuscation depends on the curvature of the obfuscation cost function $C''(\eta_H)$. If the cost of obfuscation is convex ($C''(\eta_t) > 0$), then there is an interior solution for the high-type website's optimal level of obfuscation, where $0 < \eta_H^* \leqslant \tilde{\eta}/[1-\theta]$. If the cost of obfuscation is linear ($C''(\eta_t) = 0$), then there is a corner solution for the high-type website's optimal level of obfuscation, where it does not obfuscate ($\eta_H^* = 0$).

We find that a low-type website does not obfuscate. This is intuitive, because obfuscation is costly, and it only reduces the website's demand if it is a low-type. Where the cost of obfuscation is convex (obfuscation becomes increasing difficult and costly as the website obfuscates more), the high-type website obfuscates enough that for users with high level of IT illiteracy, it is not optimal to expend effort to discover the website type. As previously explained, this occurs only for high levels of obfuscation ($E(\eta) > \tilde{\eta}$, Figure 4.(b)). Where the cost of obfuscation is linear or slightly concave, if the slope of the function is high, then the high-type website does not obfuscate. If the slope of the function is low, however, then a high-type website obfuscates similar to the previous case. Where the cost of obfuscation is strongly concave, if the slope of the function is high, then a high-type website does not obfuscate, as it is too costly to do so. If the slope of the function is low, then a high-type website obfuscates and the obfuscation level is in a way that even users with very high level of IT illiteracy ($r > \tilde{r}$) expend effort to discover the website type.

Next, we use comparative statics to study the impact of content sensitivity on obfuscation.

Proposition 1. **Content Sensitivity and Obfuscation**

(a) A low-type website does not obfuscate, irrespective of its content sensitivity.

(b) If the cost of obfuscation is linear and the content sensitivity is high ($s \geqslant \frac{1}{2C'(\frac{\tilde{\eta}}{1-\theta})[\overline{v}-\underline{v}]}$), then a high-type website does not obfuscate ($\eta_H^* = 0$). Otherwise, a high-type website obfuscates ($\eta_H^* > 0$) and this level of obfuscation is decreasing in its content sensitivity ($\partial \eta_H^* / \partial s < 0$).

To consider the impact of content sensitivity on obfuscation, note that the low-type website's demand is composed of users who use the website without expending effort, and users who expend effort. The high-type website's demand, on the other hand, is composed only of users who use the website without expending effort, because if users expend effort and discover the true level of information sharing to be high, then they do not use the website. For high-type websites, obfuscation increases demand by increasing the ratio of users who do not expend effort. High content sensitivity, however, makes it more valuable for users to discover the website type, thereby mitigating the effect of obfuscation on demand ($\partial^2 D_t/\partial\eta_t\partial s < 0$). Therefore, as content sensitivity increases, obfuscation yields diminishing benefits, resulting in website choosing a lower level of obfuscation. Particularly, where cost of obfuscation is linear or concave, if content sensitivity is sufficiently high, the website does not obfuscate.

Proposition 2. **Website Value and Obfuscation**

(a) A low-type website does not obfuscate, irrespective of the website value.

(b) The high-type website's level of obfuscation is increasing in the website value for all $X < \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})E(m)}{\theta[m_H-m_L]}$*. Increasing* X *beyond* $\frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})E(m)}{\theta[m_H-m_L]}$ *does not impact the level of obfuscation.*

As discussed above, obfuscation increases the demand of high-type websites by increasing the proportion of users who do not expend effort. While $X < \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})E(m)}{\theta[m_H-m_L]}$, that is $\eta_H^* > \frac{\tilde{\eta}}{1-\theta}$ (Figure 5(d)), as the website value increases the proportion of users whom their decisions are affected by obfuscation increases ($\tilde{r}$ increases). This intensifies the effect of obfuscation on demand, whereby $\partial^2 D_t/\partial\eta_t\partial X > 0$. Therefore, when $\eta_H^* > \frac{\tilde{\eta}}{1-\theta}$ as website value X increases, obfuscation yields higher benefits for a high-type website, resulting in these website choosing a higher level of obfuscation. When $X \geqslant \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})E(m)}{\theta[m_H-m_L]}$, that is $\eta_H^* \leqslant \frac{\tilde{\eta}}{1-\theta}$ (Figure 5(c)), the proportion of users whom their decisions are affected by obfuscation is independent of the website value. Therefore, when $\eta_H^* \leqslant \frac{\tilde{\eta}}{1-\theta}$ increasing the website value does not affect the impact of obfuscation on the website's profit.

4. Empirical Analysis

In this section, we present our empirical analysis of the website obfuscation of third-party sharing. This analysis provides some empirical support for our analytical results. We examine websites' reaction to a form of explicit user privacy request, known as Do Not Track (DNT), to analyze the level of obfuscation. Users utilize passive privacy tools, which signal to the website that such users are either concerned about their privacy, concerned about being tracked, or both. Khatibloo et al. (2018) reveal that 25% of American adults use the Do Not Track browser option to protect their privacy. Although there is generally no legal requirement in Canada (where our study is conducted) to honor a user's DNT request, it does send a clear message about the user request for privacy.

A website has three possible reactions to a DNT request. Two of these options are straightforward: it can either respect the request by reducing the number of third-parties or abuse the request by increasing the number of third-parties employed. The third option is to both add and drop third-parties. This option creates confusion in that it confuses privacy and third-party monitoring services about whether the website is complying with the DNT request or not. This is especially true for privacy-violating third-parties, i.e. those third-parties that are known to track users and collect their information. We use this latter website reaction as a measure obfuscation, as it increases the effort needed for users to determine the behavior of the website.

4.1. Data and Measurements

Alexa Internet[4] provides rankings for publisher websites within different website subject categories. From these categories, we take the list of the 500 most-visited websites from two categories: News and Health. These two categories were selected with the intention of finding website subject categories for which users have different privacy concerns and intentions to disclose personal information and browsing behavior due to the nature of the subject content. Health information is generally thought to have higher content sensitivity

[4] Alexa.com

than News. Many websites allow users to create accounts and profiles. This is particularly important for Health websites, for example `webmd.com`, `mayoclinic.org`, `medscape.com`, `drugs.com`, `psychologytoday.com`. The third-parties used by these website have access to the detailed health-related information of users. The Electronic Frontier Foundation provides some evidence for this, where a Health website shared users' health related data with third-parties (Quintin 2015). This makes the third-party sharing behavior of Health website much more sensitive than News websites. Even though not all websites collect user-specific data, user behavior on Health websites can still be sensitive, as it can be used to predict health-related facts about a user. For example, consider a user looking at a heart condition page on `cdc.gov` or `heart.org`. Such a user can be predicted to be of high risk for heart conditions by third-party trackers. Use of third-party cookies enables trackers to collect such information and depict a profile of the user with the information on their behavior in other websites. Therefore, compared to News websites, Health websites are more sensitive in terms of privacy concern. This line of reasoning is consistent with previous studies on website categories and content sensitivity, including Gopal et al. (2018).

After removing duplicate websites, the list of the 500 most-visited websites is reduced to 480 News and 441 Health websites. We also removed websites for which the average number of monthly users was unknown, resulting in a total of 676 websites (391 News and 285 Health). These were removed to limit our analysis to websites for which full data is available. We then narrowed our set of websites to a randomly selected balanced list of 200 from each category for our study to make data collection more manageable. We used Lightbeam (an add-on for Firefox) to record the connections between third-parties and websites. An automated browser accessed the homepage of each website for a particular category (that is, News or Health) with 11 runs through the entire list of websites within each category. For each of the 11 runs, the list of websites was randomly sequenced. All cookies and cached data were cleared before the first of 11 runs for each category. We removed the first run from our analysis, as this is primarily when cookies are initialized (Gopal et al. 2018). For each website, we collected HTML source code and all connections made between the website and third-parties within the first 5 seconds of visiting the website. The same process was repeated with the DNT request turned on within the browser settings.

We distinguish two types of connections between websites and third-parties. The first type is connections that are requested in the page's HTML source code, which we label as primary connections. The second type are the ones that have not been requested in the page's HTML source code, but which are redirected to other third-parties. We term such connections as secondary. We observed instances in which a particular third-party served as a primary third-party in one user visit, and a secondary third-party in another visit for the same website. We also separated third-parties based on whether they collect user information and track users' profiles. This is based on the list of tracking third-parties from EasyPrivacy[5], which is a list of third-parties that track user activity online.

We measure content sensitivity by websites category (Health has high content sensitivity and News has low content sensitivity) and use Alexa Internet ranking as a proxy for website value. Our main independent variables are $Category \in \{0,1\}$ which denotes the website category, 0 for News and 1 for Health, and $Rank$ which denotes the Alexa Internet ranking for websites. Other than website content sensitivity and value, several factors in the website environment such as website's business model and required functionalities may affect their information sharing behavior. We use two variables to control for the effect

[5] https://easylist.to/

of these factors. First, we consider whether the website uses advertising as a source of income to be a control variable. This is a binary variable denoted as 1 if the website uses advertising, labelled as Advertising. Second, we consider whether the website utilizes a user login option as a control variable. This is another binary variable denoted as 1 if the website has a login option, labelled as Login. Both variables were manually determined for each website. Because the variables Advertising (whether there is an advertising banner on the page) and Login (whether there is a login button on the page) are objectively observable, only one rater was utilized to determine the data values.

We use a multi-stage model that accounts for other managerial and economic factors. We only consider privacy-violating third-parties, as these are expected to have the negative privacy costs as perceived by users. Additionally, because websites do not have full control on the number of secondary third-parties, we only consider primary third-parties in our analysis. The websites in our study used 17,965 primary third-parties in total, where 58% of these third-parties are privacy-violating. A summary of the empirical model notation is shown in Table 3. Descriptive statistics and the and correlation matrix are provided in Tables 4 and 5, respectively.

4.2. Estimation Approach and Econometric Considerations

We examine the reaction of websites to an explicit user privacy request through DNT, to capture the websites' level of obfuscation. In response to DNT, websites have three options: respect, abuse without obfuscation, and abuse with obfuscation. Figure 6 shows the websites decision process, regarding their reaction to DNT. The decision process consists of two stages. The first stage is the choice between abusing and not abusing (respecting) the user privacy request. The second stage is a choice between obfuscate and not obfuscate for those websites who have chosen to abuse user privacy request in the first stage. To analyze this process, we use the sequential logit model proposed by Mare (1980).

Table 3 Empirical Model Notation

Notation	Definition
$Category$	A proxy for website content sensitivity, where $Category \in \{0,1\}$, 0 for News and 1 for Health (this captures s in the analytical model).
$Rank$	A proxy for website value (this captures X in the analytical model).
$Advertising$	Whether the website displays advertisements, where $Advertising \in \{0,1\}$ denoted as 1 if the website uses advertising, and 0 otherwise.
$Login$	Whether the website utilizes a user login option, where $Login \in \{0,1\}$ denoted as 1 if the website has a login option, and 0 otherwise.
$Primary$	Number of privacy-violating primary third-parties connected to the website.
$Size$	Average number of monthly users.
$Abuse$	Website decision in response to user privacy request, where $Abuse \in \{0,1\}$ denoted as 1 if the website decides to increase the number of third-parties in response to the user privacy request, and 0 otherwise.
$Abuse - Level$	Number of third-parties added in response to user privacy request.
$Obfuscate$	Website decision with respect to the obfuscation of information sharing with third-party, where $Obfuscate \in \{0,1\}$ denoted as 1 if the website decides to obfuscate, and 0 otherwise (this captures whether η_t equals zero or is greater than zero in the analytical model).
$Obfuscation - Level$	Website level of obfuscation (this captures η_t in the analytical model).

Table 4 Descriptive Statistics

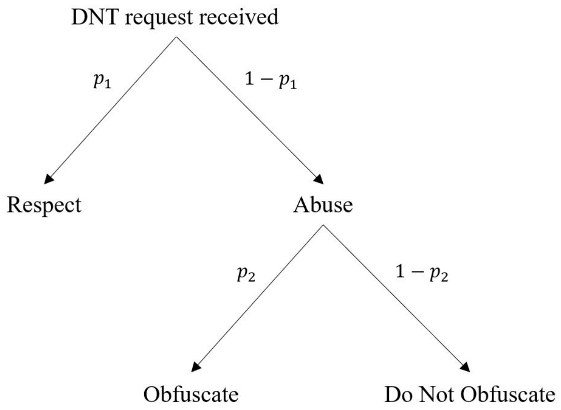
Statistic	N	Mean	St. Dev.	Min	First Quartile	Third Quartile	Max
Category	400	0.500	0.501	0	0	1	1
Rank	400	193.062	119.130	2	90.8	284.2	434
Advertising	400	0.557	0.497	0	0	1	1
Login	400	0.728	0.446	0	0	1	1
Primary	400	9.392	5.335	0	5.8	13	29
Abuse	400	0.232	0.423	0	0	0	1
Abuse-Level	400	0.355	0.872	0	0	0	9
Obfuscate	400	0.055	0.228	0	0	0	1
Obfuscation-Level	400	0.458	1.047	0	0	0	8

The sequential logit model predicts the probability of website's decisions in each stage by estimating a logistic regression. In the first stage, the sequential logit model estimates $p_1 = P(Abuse|X)$, using the overall dataset, where X is the vector of explanatory variables. In the second stage, the sequential logit model estimates $p_2 = P(Obfuscate|X, Abuse)$, using the sub-sample of websites that did not chose to respect user privacy request in the first stage. Let J be the set of all websites, and $j \in J$ represents each website. The logit

Table 5 Correlation Matrix

	Category	Rank	Advertising	Login	Primary	Abuse	Abuse-Level	Obfuscate	Obfuscation-Level
Category	1	-0.102	-0.680	-0.275	-0.521	-0.302	-0.258	-0.197	-0.294
Rank	-0.102	1	0.060	0.107	-0.005	0.006	-0.027	-0.120	-0.077
Advertising	-0.680	0.060	1	0.291	0.524	0.324	0.271	0.215	0.313
Login	-0.275	0.107	0.291	1	0.334	0.137	0.114	0.049	0.112
Primary	-0.521	-0.005	0.524	0.334	1	0.247	0.145	0.211	0.213
Abuse	-0.302	0.006	0.324	0.137	0.247	1	0.740	0.438	0.795
Abuse-Level	-0.258	-0.027	0.271	0.114	0.145	0.740	1	0.191	0.870
Obfuscate	-0.197	-0.120	0.215	0.049	0.211	0.438	0.191	1	0.608
Obfuscation-Level	-0.294	-0.077	0.313	0.112	0.213	0.795	0.870	0.608	1

Figure 6 Website Decision Process



regression models for the first and second stages are as follows:

$$Abuse_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 Advertising_j + \beta_4 Login_j + \epsilon_{1j} \tag{8}$$

$$Obfuscate_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 Advertising_j + \beta_4 Login_j + \epsilon_{2j} \tag{9}$$

In the first stage model, the dependent variable *Abuse* is a binary variable which takes the value 1 if the website increases the number of third-parties in response to the user privacy request in the first stage, and 0 otherwise. In the second stage model the dependent variable $Obfuscate$ is a binary variable which takes the value 1 if the website both adds and drops third-parties in the second stage, and 0 otherwise. There is a high correlation between websites' *Category* and *Advertising* $(-0.679, p < 0.001)$, thus if we use *Advertising* directly as a control variable, we expect to encounter multicollinearity issues.

To check the overall model for multicollinearity issues, we conduct the Farrar-Glauber test and the Theil test. The Farrar Chi-square values are 292.89 for the first model, and 46.70 for the second model. The Farrar Chi-squares are highly significant, implying that multicollinearity is present in both model specifications. The Theil test results for multicollinearity are 0.7084 (confirming multicollinearity) for the first model and 0.3980 (not confirming multicollinearity) for the second model. Next, we seek to locate specifically where the multicollinearity is located, and the results are presented in Table 6. We use the variance inflation factor (VIF), Farrar-Glauber F-test (Wi), and Klein test to locate the source of multicollinearity. the Farrar-Glauber F-test and Klein test both provide strong results for *Category* and *Advertising* being the source of the multicollinearity. While the VIF test does not provide proof of multicollinearity by itself for either model, the results are stronger for *Category* and *Advertising*. Collectively, we take this as evidence for the need to correct for multicollinearity between *Category* and *Advertising*.

Table 6 Test for Multicollinearity

	First Stage			Second Stage		
	VIF	Wi	Klein	*VIF*	Wi	Klein
Category	1.890	117.578	1	1.597	17.717	1
Rank	1.018	2.373	0	1.062	1.864	0
Advertising	1.900	118.872	1	1.482	14.326	1
Login	1.114	15.077	0	1.102	3.025	0

To deal with this problem, we develop a multi-stage model. First, we predict *Advertising* using the *Category* and all other control variables. Then we use the residuals from the model to predict the dependent variables for subsequent models. The first stage logistic regression model is:

$$Advertising_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 Login_j + \epsilon_{3j} \tag{10}$$

The coefficient for $Category$ is -3.408 and the effect of $Category$ on $Advertising$ is significant ($p < 0.01$). We take the residuals of this model and use it as a variable in the main model, denoted as $Ad_{res} = \epsilon_3$. This new variable captures all other variables that affect $Advertising$, except $Category$, $Rank$ and $Login$. In the next regression models, we use Ad_{res} as a control variable.

Unobservable variables include managerial, economic, and other factors, may impact the website decision regarding third-party usage. To resolve the omitted variable bias problem, we use the residuals from the following model ($Prim_{res} = \epsilon_{4j}$), which captures other factors that affect websites decisions, as another control variable in our main models. Our dependent variable ($Primary_j$) is the number of primary third-parties. Because this is a count value, we use negative binomial regression model which is appropriate for modeling count values. We also control for the total number of third-parties using an offset.

$$Primary_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 \epsilon_{3j} + \beta_4 Login_j + \epsilon_{4j} \tag{11}$$

Finally, to avoid multicollinearity issues and omitted variable bias, we estimate the following two logit regression models.

$$Abuse_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 \epsilon_{3j} + \beta_4 Login_j + \beta_5 \epsilon_{4j} + \epsilon_{5j} \tag{12}$$

$$Obfuscate_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 \epsilon_{3j} + \beta_4 Login_j + \beta_5 \epsilon_{4j} + \epsilon_{6j} \tag{13}$$

We also examine the magnitude of websites' reaction to user privacy request. We capture this using two additional dependent variables, $Abuse-Level$ and $Obfuscation-Level$. $Abuse-Level$ is the number of third-parties added in response to user privacy request, and $Obfuscation-Level$ is the summation of both the number of third-parties added and dropped with the user privacy request. Because these additional dependent variables are count values, we analyze them using negative binomial regression model.

The results for the first and second stage models are provided in Table 7. The effect of $Category$ on both $Abuse$ and $Abuse-Level$ in the first stage is significant and the coefficient is negative. This result shows that websites with less sensitive content are more likely to abuse user privacy request and the extent of abuse is also higher for the websites with less sensitive content. In the second stage the impact of $Category$ on $Obfuscate$ is significant and the coefficient is negative, implying that websites with low content sensitivity are more likely to obfuscate. This result is consistent with the findings in Proposition 1, where the obfuscation cost function is linear. The impact of $Category$ on $Obfuscation-Level$ is also significant in the negative binomial model and the coefficient is negative, which indicates that among the websites that obfuscate, the level of obfuscation of websites with low content sensitivity is higher. This result is consistent with the findings for both linear and convex obfuscation cost functions in Proposition 1. The impact of $Rank$ on $Obfuscate$ is significant and the coefficient is negative (-0.007), which indicates that less prominent websites are less likely to obfuscate. The impact of $Rank$ on $Obfuscation-Level$ is also negative and significant in the negative binomial model. This implies that among the websites that obfuscate, more prominent websites employ higher levels of obfuscation. These findings are in line with Proposition 2.

4.3. Robustness Check

To further investigate the robustness of our results, we repeat our analysis using both Primary and Secondary (Total) third-parties. The results of this analysis (as shown in Table 8) are fairly consistent with our previous findings. The impact of $Rank$ on $Obfuscate$ is no longer significant, as the decision to share information with secondary third-parties is not directly made by the website.

We also employed the average number of monthly users ($Size$) to reflect $Rank$ as a robustness check. We replaced $Rank$ with $log(Size)$ in all models. Results are provided in

Table 7 Website Reaction to DNT (Two-stage Model) Using Primary Third-Parties

	First Stage		Second Stage	
	$Abuse$	$Abuse-Level$	$Obfuscate$	$Obfuscation-Level$
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
$Category$	−1.542***	−1.530***	−2.656*	−0.585***
	(0.290)	(0.272)	(1.241)	(0.217)
$Rank$	0.000	−0.001	−0.007***	−0.001***
	(0.001)	(0.001)	(0.002)	(0.000)
Ad_{res}	0.453***	0.421***	−0.987	0.141
	(0.144)	(0.135)	(0.673)	(0.096)
$Login$	0.398	0.500	0.048	0.091
	(0.336)	(0.316)	(0.807)	(0.217)
$Prim_{res}$	0.327	−0.100	1.336	−0.241
	(0.344)	(0.319)	(1.177)	(0.250)
Constant	−0.736*	−0.756**	0.195	0.946***
	(0.394)	(0.365)	(0.979)	(0.231)
Observations	400	400	93	93
AIC	393.349	570.169	94.972	294.481
Hosmer-Lemeshow p-value	0.599	0	0.439	1.000
McFadden's R^2	0.121	0.082	0.184	0.051
Cox & Snell R^2	0.123	0.117	0.183	0.151
Cameron & Windmeijer R^2	0.121	0.191	0.184	0.225

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table 8 Website Reaction to DNT (Two-stage Model) Using Total (Primary+Secondary) Third-Parties

	First Stage		Second Stage	
	$Abuse$	$Abuse-Level$	$Obfuscate$	$Obfuscation-Level$
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
$Category$	-0.674***	-1.512***	-2.591***	-1.305***
	(0.246)	(0.275)	(0.599)	(0.145)
$Rank$	0.001	-0.001	-0.001	-0.001**
	(0.001)	(0.001)	(0.002)	(0.0005)
Ad_{res}	0.504***	0.563***	1.111***	0.452***
	(0.128)	(0.133)	(0.299)	(0.074)
$Login$	0.282	0.451	0.116	0.116
	(0.287)	(0.321)	(0.614)	(0.165)
$Total_{res}$	-0.024	0.586*	3.597***	1.149***
	(0.288)	(0.309)	(0.857)	(0.158)
Constant	-1.079***	0.761*	2.692***	3.007***
	(0.354)	(0.392)	(0.841)	(0.202)
Observations	400	400	93	93
AIC	462.831	1045.172	102.489	747.015
Hosmer-Lemeshow p-value	0.069	1.000	0.016	1.000
McFadden's R^2	0.064	0.042	0.377	0.113
Cox & Snell R^2	0.074	0.108	0.376	0.552
Cameron & Windmeijer R^2	0.064	0.171	0.377	0.537

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table 9 (for primary third-parties) and Table 10 (for total third-parties). It can be seen that both tables are consistent with the main results.

Table 9 Website Reaction to DNT (Two-stage Model With log(Size)) Using Primary Third-Parties

	First Stage		Second Stage	
	$Abuse$	$Abuse - Level$	$Obfuscate$	$Obfuscation - Level$
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
$Category$	-1.442***	-1.387***	-2.178*	-0.453**
	(0.291)	(0.272)	(1.233)	(0.215)
$log(Size)$	0.190**	0.209***	0.589***	0.144***
	(0.088)	(0.075)	(0.185)	(0.046)
Ad_{res}	0.442***	0.407***	0.949	0.129
	(0.144)	(0.134)	(0.677)	(0.097)
$Login$	0.381	0.459	-0.166	0.039
	(0.339)	(0.314)	(0.834)	(0.216)
$Prim_{res}$	0.289	-0.023	1.518	-0.159
	(0.349)	(0.318)	(0.977)	(0.252)
Constant	-3.457***	-3.836***	-9.064***	-1.309*
	(1.237)	(1.073)	(2.747)	(0.681)
Observations	400	400	93	93
AIC	389.616	564.693	94.535	292.251
Hosmer-Lemeshow p-value	0.494	1.000	0.652	1.000
McFadden's R^2	0.129	0.092	0.189	0.059
Cox & Snell R^2	0.131	0.129	0.187	0.171
Cameron & Windmeijer R^2	0.129	0.210	0.188	0.259

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table 10 Website Reaction to DNT (Two-stage Model With log(Size)) Using Total Third-Parties

	First Stage		Second Stage	
	$Abuse$	$Abuse - Level$	$Obfuscate$	$Obfuscation - Level$
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
$Category$	-0.715***	-1.452***	-2.506***	-1.252***
	(0.248)	(0.277)	(0.594)	(0.143)
$log(Size)$	-0.040	0.128	0.193	0.118***
	(0.084)	(0.094)	(0.204)	(0.044)
Ad_{res}	0.505***	0.560***	1.110***	0.448***
	(0.127)	(0.132)	(0.296)	(0.074)
$Login$	0.308	0.351	-0.034	0.038
	(0.286)	(0.318)	(0.652)	(0.168)
$Total_{res}$	-0.017	0.577*	3.490***	1.147***
	(0.286)	(0.309)	(0.847)	(0.155)
Constant	-0.327	-1.035	0.054	1.276**
	(1.161)	(1.307)	(2.612)	(0.580)
Observations	400	400	93	93
AIC	463.924	102.555	1043.890	745.036
Hosmer-Lemeshow p-value	0.675	0.020	1.000	1.000
McFadden's R^2	0.062	0.376	0.044	0.115
Cox & Snell R^2	0.072	0.376	0.111	0.560
Cameron & Windmeijer R^2	0.062	0.376	0.176	0.544

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Moreover, we analyze the empirical results for estimating coefficients in equations (8) and (9) directly. Results of the one-stage models are provided in Table 11. The effect of $Category$ on $Abuse$ and $Obfuscate$ are significant, consistent with the main results (two-stage models). However the effect of $Category$ on abuse and obfuscation levels are not significant. The effect of $Rank$ on $Obfuscation-Level$ is significant, consistent with the main results. However the effect of $Rank$ on $Obfuscate$ is not significant. While some of the one-stage model results are different from two-stage models, the results for the main coefficients are consistent with the analytical model findings in Proposition 1 and Proposition 2.

Table 11 Website Reaction to DNT (One-stage Model) Using Primary Third-Parties

	First Stage		Second Stage			
	Abuse	*Abuse − Level*	*Obfuscate*	*Obfuscation − Level*		
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)		
Category			-0.754**	-1.121	-0.744**	-0.279
			(0.346)	(0.891)	(0.322)	(0.232)
Rank			-0.001	-0.007***	-0.001	-0.002***
			(0.001)	(0.002)	(0.001)	(0.001)
Advertising			1.300***	17.353	1.305***	0.519*
			(0.386)	(1587.729)	(0.369)	(0.294)
Login			0.269	-0.116	0.325	0.033
			(0.337)	(0.741)	(0.315)	(0.211)
Constant			-1.866***	-16.860	-1.782***	0.552
			(0.504)	(1587.730)	(0.473)	(0.368)
Observations			400	400	93	93
AIC			391.651	566.100	92.946	292.385
Hosmer-Lemeshow p-value			0.537	0	0.718	1.000
McFadden's R^2			0.120	0.086	0.184	0.052
Cox & Snell R^2			0.122	0.122	0.183	0.152
Cameron & Windmeijer R^2			0.120	0.198	0.185	0.227

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

5. Concluding Remarks

With the increasing importance of user privacy, understanding the interaction of users, websites, and third-parties is more important than ever. In this paper, we focus on the implicit contract between users and third-parties with respect to how websites share user information, as well as their effort to obfuscate the use of third parties. Developing analytical models for theoretical development and empirical models for external validation, this study examines the websites' obfuscation of user information sharing with third-parties. Our analytical model examines the website's decision regarding the extent of obfuscation of information sharing, given user privacy concerns. Interestingly and counter-intuitively, websites reduce the level of obfuscation as user privacy concern increases. Moreover, prominent websites are more prone to obfuscate the sharing of user information with third-parties.

Our results have important managerial insights for websites concerned about user privacy and informed user consent. Obfuscation of information sharing impairs the user's ability to make informed consent with respect to the privacy risks of using a website. An important contribution of this work is to explain the websites' incentive to obfuscate user information sharing. This has several implications for website practices toward information sharing. First, if user privacy is substantially impacted by the sensitive nature of the website content, websites should refrain from excessive monetization and the temptation to obfuscate third-party sharing. Secondly, it is primarily the most popular websites have the latitude to obfuscate. Managers of relatively less popular websites with more sensitive content have a greater incentive to refrain from obfuscation. Going beyond our results, we see current examples of heads of large platforms being called before the U.S. Congress to explain abusive practices. Eventually, threats of anti-trust enforcement and reputational damage to the brand could exact a substantial penalty for what are likely to be the comparatively moderate benefits of obfuscation.

Additionally, our results have important implications for policy makers concerned with website exploitation of user information. Policy makers need to be aware of information asymmetry problems and user privacy implications of information sharing as they design policies. One area of active policy making is in the domain of data protection regulation, for example European Union's General Data Protection Regulation (GDPR) or California's Consumer Privacy Act (CCPA). GDPR and CCPA require websites to acquire user consent (allow the user to either opt-in or opt-out) before their information is shared with third-parties. Some may argue that such regulation might make the issue of obfuscation moot.

However, this is not necessarily the case. For the same reasons that website privacy policies do not convey all the information to users, acquiring consent does not necessarily resolve the information asymmetry between websites and users. First, the list of third-parties that websites are mandated to provide to users are typically long, confusing, and ineffective. Having such a list does not resolve information asymmetry as users have to engage in additional effort regarding the source and credibility of third-parties. Similar to privacy policies, these provided lists are often designed and presented in ways to confuse users and can persuade them to make risky privacy choices.

To further investigate the impact of data protection policies on obfuscation, we examined third-parties from a small sample of websites under the GDPR legal environment when accessing using a European IP address. We collected information from the reported lists of third-party vendors provided by websites (required by GDPR), examined the HTML source code for links to third-parties, and collected information on third-parties with whom the websites communicated using Lightbeam. We find that websites continue to use secondary third-parties under GDPR. It is difficult to determine if websites properly report these third-parties in the provided list because the list only consists of vendors' company names, not the URLs. This makes it hard for users to accurately determine the third-parties that have access to their data. We also examined the sharing behavior with and without DNT and find that some websites add and/or drop third-parties from their reported list of third-parties in reaction to the DNT request. These observations imply that obfuscation continues to be present even under the GDPR legal regime.

Thus, we are skeptical that current GDPR legislation, in and of itself, can effectively solve the problem of obfuscation. The cat and mouse game between regulators and obfuscators is likely to continue with ever evolving technology. Perhaps standardized and required transparency is the regulator's only real tool. Standardized disclosures are required for a variety of consumer financial contracts and corporate financial statement disclosures, and the same concepts could be applied to the website level of information sharing. Fines could reasonably be imposed for disclosure violations.

Websites seem to be reacting to the market forces surrounding user privacy concern and their ability or inability to exploit users' lack of information through third-party obfuscation. We confirm that it is primarily the most prominent websites that obfuscate their use of third-parties. Our analysis illustrates that obfuscation is not simply a matter of ability related to size and resources, but rather is a form of strategic exploitation. It is the large and prominent websites that obfuscate in order to additionally monetize user information. Small websites could use these insights to inform their interaction with users in terms of privacy and obfuscation. Even though we assume that small websites would exploit the if they could, that should not prevent small websites from highlighting how honest they are compared to their more dominant competitors.

We provide a simple and elegant model to capture the nuances involved in the obfuscation of third-parties. That said, in reality, the level of information sharing is more likely determined at the same time as the level of obfuscation in a dynamic manner. This makes a dynamic model of incomplete information a useful avenue for future research on this topic. Future work would also benefit from more refined measures of content sensitivity. If we have a better measure for content sensitivity, then researchers would be able to utilize this robust measure of content sensitivity to expand the number of categories and further validate our findings. Much future work is possible regarding the measurement of content sensitivity.

Another area for future work includes connecting this research to the impact of search engine optimization on obfuscation and user privacy. It would be useful to examine the effect of competition on websites' strategic decision-making with respect to obfuscation and privacy violation. Additionally, future analysis could examine the use of secondary third-parties, such as third-party advertisers not directly called by the website, using a sequential principal-agent model. Even though our study provides important implications for policy-making, we do not analytically address the policy-makers' incentive to limit or prohibit the level of obfuscation that websites utilize. Such analysis can have provide additional insights to help with policy-making in future studies.

References

Acar G, Englehardt S, Narayanan A (2020) No boundaries: data exfiltration by third parties embedded on web pages. *Proceedings on Privacy Enhancing Technologies* 2020(4):220–238.

Akerlof G (1970) The market for lemons: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics* 84(3):488–500.

Akerlof GA, Shiller RJ (2015) *Phishing for phools: The economics of manipulation and deception* (Princeton University Press).

Akerlof GA, Shiller RJ (2016) Manipulation and deception as part of a phishing equilibrium. *Business Economics* 51(4):207–212.

Arrow KJ (1985) The economics of agency. Technical report, In J. W. Pratt R. J. Zeckhauser (Eds.), Principals and agents: The structure of business (pp. 37-51). Boston: Harvard Business School Press.

Bai X, Gopal R, Nunez M, Zhdanov D (2012) On the prevention of fraud and privacy exposure in process information flow. *INFORMS Journal on Computing* 24(3):416–432.

Bushee BJ, Gow ID, Taylor DJ (2018) Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research* 56(1):85–121.

Chellappa RK, Shivendu S (2007) An economic model of privacy: A property rights approach to regulatory choices for online personalization. *Journal of Management Information Systems* 24(3):193–225.

Cisco (2021) Cisco 2021 consumer privacy survey. `https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/cisco-cybersecurity-series-2021-cps.pdf`, accessed: 2021-11-20.

Eisenhardt KM (1989) Agency theory: An assessment and review. *Academy of management review* 14(1):57–74.

Ellison G, Wolitzky A (2012) A search cost model of obfuscation. *The RAND Journal of Economics* 43(3):417–441.

Englehardt S, Narayanan A (2016) Online tracking: A 1-million-site measurement and analysis. *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 1388–1401.

Englehardt S, Reisman D, Eubank C, Zimmerman P, Mayer J, Narayanan A, Felten EW (2015) Cookies that give you away: The surveillance implications of web tracking. *Proceedings of the 24th International Conference on World Wide Web*, 289–299.

Goldstein A, Eaton C (2021) Asymmetry by design? identity obfuscation, reputational pressure, and consumer predation in us for-profit higher education. *American Sociological Review* 86(5):896–933.

Gopal RD, Hidaji H, Patterson RA, Rolland E, Zhdanov D (2018) How much to share with third parties? user privacy concerns and website dilemmas. *MIS Quarterly* 42(1):143–164.

Gu Y, Wenzel T (2014) Strategic obfuscation and consumer protection policy. *The Journal of Industrial Economics* 62(4):632–660.

Gupta R, Bagchi A, Sarkar S (2007) Improving linkage of web pages. *INFORMS Journal on Computing* 19(1):127–136.

Harsanyi JC (1968) Games with incomplete information played by "bayesian" players part ii. bayesian equilibrium points. *Management Science* 14(5):320–334.

Hölmstrom B (1979) Moral hazard and observability. *The Bell journal of economics* 74–91.

Jenson MC, Meckling WH (1976) Theory of the firm: managerial behavior, agency costs and ownership structure. *Journal of financial economics* 3(4):305–360.

Khatibloo F, Liu S, Pilecki M, Flug M, Hartig K (2018) Right your privacy ship before it capsizes. *Forrester* URL `http://www.forrester.com/report/Right+Your+Privacy+Ship+Before+It+Capsizes/-/E-RES133381`.

Krishnamurthy B, Naryshkin K, Wills C (2011) Privacy leakage vs. protection measures: the growing disconnect. *Proceedings of the Web*, volume 2, 1–10.

Li Y (2012) Theories in online information privacy research: A critical review and an integrated framework. *Decision Support Systems* 54(1):471–481.

Libert T (2015) Exposing the hidden web: An analysis of third-party http requests on 1 million websites. *arXiv preprint arXiv:1511.00619* .

Mare RD (1980) Social background and school continuation decisions. *Journal of the american statistical association* 75(370):295–305.

Pavlou PA (2011) State of the information privacy literature: Where are we now and where should we go? *MIS Quarterly* 977–988.

PewResearch (2019) Americans and privacy: Concerned, confused and feeling lack of control over their personal information. https://www.pewresearch.org/internet/2019/11/15/americans-attitudes-and-experiences-with-privacy-policies-and-laws, accessed: 2021-11-24.

Quintin C (2015) Healthcare. gov sends personal data to dozens of tracking websites. *Electronic Frontier Foundation* .

Roesner F, Kohno T, Wetherall D (2012) Detecting and defending against third-party tracking on the web. *Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*, 155–168.

Salesforce (2019) State of the connected customer. https://www.salesforce.com/news/stories/state-of-the-connected-customer-report-outlines-changing-standards-for-customer-engagement/, accessed: 2021-11-20.

Smith HJ, Dinev T, Xu H (2011) Information privacy research: an interdisciplinary review. *MIS Quarterly* 35(4):989–1016.

TechTimes (2020) Safe browsing made easy: Benefits of using web of trust. https://www.techtimes.com/articles/248191/20200319/safe-browsing-made-easy-benefits-of-using-web-of-trust, accessed: 2021-11-24.

Vincent J (2021) California bans 'dark patterns' that trick users into giving away their personal data. *The Verge,* `https://www.theverge.com/2021/3/16/22333506/california-bans-dark-patterns-opt-out-selling-data` .

Wilson CM (2010) Ordered search and equilibrium obfuscation. *International Journal of Industrial Organization* 28(5):496–506.

Yang Y, Liu H, Cai Y (2013) Discovery of online shopping patterns across websites. *INFORMS Journal on Computing* 25(1):161–176.

Yu Z, Macbeth S, Modi K, Pujol JM (2016) Tracking the trackers. *Proceedings of the 25th International Conference on World Wide Web*, 121–132.

Now You See it, Now You Don't: Obfuscation of Online Third-Party Information Sharing (Online Supplement)

Ashkan Eshghi
Haskayne School of Business, University of Calgary, Calgary, Alberta T2N 1N4, Canada, ashkan.eshghi1@ucalgary.ca

Ram Gopal
Warwick Business School, University of Warwick, ram.gopal@wbs.ac.uk

Hooman Hidaji, Raymond Patterson
Haskayne School of Business, University of Calgary, Calgary, Alberta T2N 1N4, Canada,
hooman.hidaji@haskayne.ucalgary.ca, raymond.patterson@ucalgary.ca

Proof of Lemma 1

We first describe the details for deriving the equilibrium. As discussed, The Bayesian Nash equilibria of this game are vectors $(\eta_t^*, K_{v,r}^*)$ such that:

$$\begin{aligned} B_{v,r}(\eta_L^*, \eta_H^*) &= K_{v,r}^* \quad , \qquad \forall \ (v \in [\underline{v}, \overline{v}] \,,\, r \in [0,1]) \\ B_t(\{K_{v,r}^* | \ \underline{v} \leqslant v \leqslant \overline{v} \,,\, 0 \leqslant r \leqslant 1\}) &= \eta_t^* \quad , \qquad \forall \ t \in \{L, H\} \end{aligned} \tag{1}$$

We first focus on the low-type website's decision. Demand for a low-type website given the best response function of the users and the high-type website is $\frac{\int_0^{\min\{\tilde{r},1\}} \hat{V}\, dr + \int_{\min\{\tilde{r},1\}}^1 \tilde{v}\, dr - \underline{v}}{\overline{v} - \underline{v}}$.

We can re-write this piecewise function as:

$$D_L(\eta_L, \eta_H) = \begin{cases} \frac{2\theta[X - s m_L \underline{v}] - [\theta\eta_L + [1-\theta]\eta_H]}{2 s m_L \theta[\overline{v} - \underline{v}]} & \text{if } \eta_L \leqslant \frac{\tilde{\eta} - [1-\theta]\eta_H}{\theta} \\ \frac{\frac{[m_H - m_L]^2 [1-\theta]^2 \theta X^2 - 2 X m_L [\theta\eta_L + [1-\theta]\eta_H][\theta m_L + [1-\theta] m_H]}{2 s m_L [\theta\eta_L + [1-\theta]\eta_H][\theta m_L + [1-\theta] m_H]^2} - \underline{v}}{\overline{v} - \underline{v}} & \text{if } \eta_L > \frac{\tilde{\eta} - [1-\theta]\eta_H}{\theta} \end{cases} \tag{2}$$

We can calculate $\frac{\partial D_L}{\partial \eta_L}$ as:

$$\frac{\partial D_L}{\partial \eta_L} = \begin{cases} -\frac{1}{2sm_L[\overline{v}-\underline{v}]} < 0 & \text{if } \eta_L \leqslant \frac{\tilde{\eta}-[1-\theta]\eta_H}{\theta} \\ -\frac{[m_H-m_L]^2[1-\theta]^2\theta^2X^2}{2sm_L[\overline{v}-\underline{v}][\theta\eta_L+[1-\theta]\eta_H]^2[\theta m_L+[1-\theta]m_H]^2} < 0 & \text{if } \eta_L > \frac{\tilde{\eta}-[1-\theta]\eta_H}{\theta} \end{cases} \tag{3}$$

Therefore $\frac{\partial D_L}{\partial \eta_L} < 0$. The profit of a low-type website is given as $\Pi_L = D_L(\eta_L, \eta_H)m_L - C(\eta_L)$, thus, $\frac{\partial \Pi_L}{\partial \eta_L} = \frac{\partial D_L}{\partial \eta_L}m_L - C'(\eta_L)$. We know that $C'(\eta_L) > 0$, therefore $\frac{\partial \Pi_L}{\partial \eta_L} < 0$. For a low-type website, there is no incentive to obfuscate, thus $\eta_L^* = 0$.

Now we focus on a high-type website's decision. Demand for a high-type website given the best response function of the users is $\frac{\int_0^{\min\{\tilde{r},1\}} \check{V}\,dr + \int_{\min\{\tilde{r},1\}}^1 \tilde{v}\,dr - \underline{v}}{\overline{v}-\underline{v}}$. We can rewrite this given $\eta_L^* = 0$ as:

$$D_H(\eta_L = 0, \eta_H) = \begin{cases} \frac{\eta_H + 2X - 2sm_H\underline{v}}{2sm_H[\overline{v}-\underline{v}]} & \text{if } \eta_H \leqslant \frac{\tilde{\eta}}{1-\theta} \\ \frac{\frac{[m_H-m_L]^2\theta^2X + 2\eta_H m_H[\theta m_L+[1-\theta]m_H]}{2s[\theta\eta_L+[1-\theta]\eta_H]^2} - \underline{v}}{\overline{v}-\underline{v}} & \text{if } \eta_H > \frac{\tilde{\eta}}{1-\theta} \end{cases} \tag{4}$$

We can calculate $\frac{\partial D_H}{\partial \eta_H}$ as:

$$\frac{\partial D_H}{\partial \eta_H} = \begin{cases} \frac{1}{2sm_H[\overline{v}-\underline{v}]} > 0 & \text{if } \eta_H \leqslant \frac{\tilde{\eta}}{1-\theta} \\ \frac{[m_H-m_L]^2\theta^2X^2}{2sm_H[\overline{v}-\underline{v}]\eta_H^2[\theta m_L+[1-\theta]m_H]^2} > 0 & \text{if } \eta_H > \frac{\tilde{\eta}}{1-\theta} \end{cases} \tag{5}$$

Therefore $\frac{\partial D_H}{\partial \eta_H} > 0$. The profit of a high-type website is given as $\Pi_H = D_H(\eta_L, \eta_H)m_H - C(\eta_H)$, thus, $\frac{\partial \Pi_H}{\partial \eta_H} = \frac{\partial D_H}{\partial \eta_H}m_H - C'(\eta_H)$. To find the optimal level of obfuscation for the high-type website, we need to derive the optimal solution for each of the demand function intervals and then compare these optimal solutions. The first-order and second-order conditions are as follows:

$$\text{FOC}: \frac{\partial \Pi_H}{\partial \eta_H} = 0 \Rightarrow \begin{cases} (a) \quad \frac{1}{2s[\overline{v}-\underline{v}]} - C'(\eta_H) = 0 & \text{if } \eta_H \leqslant \frac{\tilde{\eta}}{1-\theta} \\ (b) \quad \frac{[m_H-m_L]^2\theta^2X^2}{2s[\overline{v}-\underline{v}]\eta_H^2[\theta m_L+[1-\theta]m_H]^2} - C'(\eta_H) = 0 & \text{if } \eta_H > \frac{\tilde{\eta}}{1-\theta} \end{cases} \tag{6}$$

$$\text{SOC}:\ \frac{\partial^2 \Pi_H}{\partial \eta_H^2} < 0 \Rightarrow \begin{cases} (a) \quad -C''(\eta_H) < 0 & \text{if } \eta_H \leqslant \frac{\tilde{\eta}}{1-\theta} \\ (b) \quad -\frac{[m_H-m_L]^2\theta^2 X^2}{s[\overline{v}-\underline{v}]\eta_H^3[\theta m_L+[1-\theta]m_H]^2} - C''(\eta_H) = 0 & \text{if } \eta_H > \frac{\tilde{\eta}}{1-\theta} \end{cases} \tag{7}$$

From FOC(a) in (6) and the assumption $C''(\eta_t) \geqslant 0$, we obtain that if $C'(\frac{\tilde{\eta}}{1-\theta}) < \frac{1}{2s[\overline{v}-\underline{v}]}$, then there is no optimal solution where $\eta_H^* \leqslant \frac{\tilde{\eta}}{1-\theta}$. From FOC(b) in (6), SOC(b) in (7) and the assumption $C''(\eta_t) \geqslant 0$ we obtain that if $C'(\frac{\tilde{\eta}}{1-\theta}) < \frac{1}{2s[\overline{v}-\underline{v}]}$, then there is an interior solution where $\eta_H^* > \frac{\tilde{\eta}}{1-\theta}$. From FOC(b) in (6) and the assumption $C''(\eta_t) \geqslant 0$ we obtain that if $C'(\frac{\tilde{\eta}}{1-\theta}) \geqslant \frac{1}{2s[\overline{v}-\underline{v}]}$, then there is no interior solution where $\eta_H^* > \frac{\tilde{\eta}}{1-\theta}$. From FOC(a) in (6) and SOC(a) in (7) we obtain that if $C'(\frac{\tilde{\eta}}{1-\theta}) \geqslant \frac{1}{2s[\overline{v}-\underline{v}]}$ and $C''(\eta_t) > 0$, then there is an interior solution where $\eta_H^* \leqslant \frac{\tilde{\eta}}{1-\theta}$. If $C'(\frac{\tilde{\eta}}{1-\theta}) \geqslant \frac{1}{2s[\overline{v}-\underline{v}]}$ and $C''(\eta_t) = 0$, then from SOC(a) in (7) and the assumption $C'(\eta_t) > 0$, we obtain that there is a corner solution where $\eta_H^* = 0$. Thus, the characteristics of the high-type website's level of obfuscation are as follows:

$$\begin{aligned} \eta_H^* &>, \frac{\tilde{\eta}}{1-\theta} && \text{if } \ C'(\frac{\tilde{\eta}}{1-\theta}) < \frac{1}{2s[\overline{v}-\underline{v}]} \\ \eta_H^* &\leqslant \frac{\tilde{\eta}}{1-\theta}, && \text{if } \ C'(\frac{\tilde{\eta}}{1-\theta}) \geqslant \frac{1}{2s[\overline{v}-\underline{v}]} \quad \& \quad C''(\eta) > 0 \\ \eta_H^* &= 0, && \text{if } \ C'(\frac{\tilde{\eta}}{1-\theta}) \geqslant \frac{1}{2s[\overline{v}-\underline{v}]} \quad \& \quad C''(\eta) = 0 \qquad \blacksquare \end{aligned} \tag{8}$$

Proof of Proposition 1

From Lemma 1(a) we know that $\eta_L^* = 0$. Therefore a low-type website does not obfuscate, irrespective of content sensitivity.

From Lemma 1(c) we know that if $C''(\eta_t) = 0$ and $C'(\eta_t) \geqslant \frac{1}{2s[\overline{v}-\underline{v}]}$, then $\eta_H^* = 0$. Therefore, where the cost of obfuscation is linear, a high-type website does not obfuscate if $s \geqslant \frac{1}{2C'(\eta_t)[\overline{v}-\underline{v}]}$. In other conditions, to find the impact of content sensitivity on a high-type website's level of obfuscation, we need to determine the sign of $\frac{\partial \eta_H^*}{\partial s}$ in all possible equilibrium solutions. By the implicit function theorem applied to FOC in (6), we obtain:

$$\frac{\partial \eta_H^*}{\partial s} = -\frac{\frac{\partial^2 \Pi_H}{\partial \eta_H^* \partial s}}{\frac{\partial^2 \Pi_H}{\partial {\eta_H^*}^2}} = \begin{cases} -\frac{-\frac{1}{2s^2[\overline{v}-\underline{v}]}}{\frac{\partial^2 \Pi_H}{\partial {\eta_H^*}^2}} < 0 & \text{if } 0 < \eta_H^* \leqslant \frac{\tilde{\eta}}{1-\theta} \\ -\frac{-\frac{[m_H - m_L]^2 \theta^2 X^2}{2s^2 {\eta_H^*}^2 [\theta m_L + [1-\theta] m_H]^2 [\overline{v}-\underline{v}]}}{\frac{\partial^2 \Pi_H}{\partial {\eta_H^*}^2}} < 0 & \text{if } \eta_H^* > \frac{\tilde{\eta}}{1-\theta} \end{cases} \tag{9}$$

From (9) we have $\frac{\partial \eta_H^*}{\partial s} < 0$ for all $\eta_H^* > 0$, thus, if a high-type website obfuscates, then the level of obfuscation is decreasing in content sensitivity. ■

Proof of Proposition 2

From Lemma 1(a) we know that $\eta_L^* = 0$. Therefore a low-type website does not obfuscate irrespective of website value.

From Lemma 1(c) we know that if $C''(\eta_t) = 0$ and $C'(\eta_t) \geqslant \frac{1}{2s[\overline{v}-\underline{v}]}$, then $\eta_H^* = 0$ irrespective of website value. In other conditions, to find the impact of website value on a high-type website's level of obfuscation, we need to determine the sign of $\frac{\partial \eta_H^*}{\partial X}$ in all possible interior equilibrium solutions. However, the condition threshold $(\frac{\tilde{\eta}}{1-\theta})$ in deriving equilibrium solutions depends on X. Given the assumption $C'(\eta) > 0$ we can rewrite the equilibrium solutions as:

$$\begin{aligned} \eta_H^* &> \frac{\tilde{\eta}}{1-\theta}, && \text{if} \quad X < \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})[\theta m_L + [1-\theta] m_H]}{\theta[m_H - m_L]} \\ 0 < \eta_H^* &\leqslant \frac{\tilde{\eta}}{1-\theta}, && \text{if} \quad X \geqslant \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})[\theta m_L + [1-\theta] m_H]}{\theta[m_H - m_L]} \quad \& \quad C''(\eta_t) > 0 \\ \eta_H^* &= 0, && \text{if} \quad X \geqslant \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})[\theta m_L + [1-\theta] m_H]}{\theta[m_H - m_L]} \quad \& \quad C''(\eta_t) = 0 \end{aligned} \tag{10}$$

By the implicit function theorem applied to FOC in (6), we obtain:

$$\frac{\partial \eta_H^*}{\partial X} = -\frac{\frac{\partial^2 \Pi_H}{\partial \eta_H^* \partial X}}{\frac{\partial^2 \Pi_H}{\partial {\eta_H^*}^2}} = \begin{cases} 0 & \text{if } X \geqslant \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})[\theta m_L + [1-\theta] m_H]}{\theta[m_H - m_L]} \\ -\frac{\frac{[m_H - m_L]^2 \theta^2 X}{s {\eta_H^*}^2 [\theta m_L + [1-\theta] m_H]^2 [\overline{v}-\underline{v}]}}{\frac{\partial^2 \Pi_H}{\partial {\eta_H^*}^2}} > 0 & \text{if } X < \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})[\theta m_L + [1-\theta] m_H]}{\theta[m_H - m_L]} \end{cases} \tag{11}$$

From 11 we can conclude that as far as $X < \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})[\theta m_L+[1-\theta]m_H]}{\theta[m_H-m_L]}$ the optimal level of obfuscation is increasing in X ($\frac{\partial\eta_H^*}{\partial X} > 0$). When $X \geqslant \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})[\theta m_L+[1-\theta]m_H]}{\theta[m_H-m_L]}$, changing X does not impact η_H^*. ■

INFORMS Journal on Computing

Response to Review Team's Comments on: Now You See it, Now You Don't: Obfuscation of Online Third-Party Information Sharing, JOC-2021-03-OA-070.R1

Overview of Revision

We would like to thank the review team's very valuable comments on this manuscript. These comments have helped us to continue to improve the paper. We provide our responses to the review team's comments in this document. For clarity, we provide the review team's original comments in blue, and our responses to the comments and resulting changes to the manuscript are provided in black. To aid with the review, we also attach a version of the paper with tracked changes in this new revision.

Associate Editor's Comments

The revised paper was sent back to the original review team. All three reviewers acknowledged that the authors have put significant effort in this revision. They have completed revamped the analytical model. The new model is clean, elegant and yields interesting predictions that are validated with the empirical testing. As a result, both R1 recommends minor revision, R2 recommends acceptance, and R3 recommends major revision. After reading the revised manuscript and the authors' response letter, I agree with the review team that the authors have satisfactorily addressed the major concerns from the previous round. Although R3 still has some questions about the empirical analysis, I believe the comments can be addressed with another round of careful revision. Overall, I would like to congratulate the authors for their successful revision in this round. I look forward to receiving the further improved manuscript that meets the high expectation of IJOC publication.

Thank you for your support and kind words. We carefully address all reviewer comments below.

Reviewer 1's Comments

I thank the authors for addressing my comments during the previous round. In this round, I have only one clarification. There are a variety of health-related websites. Unless they contain patient-specific data, why they should be considered sensitive? Some details about the website would be helpful; the data should be representative of the case.

Thank you for this comment. Our dataset includes many websites that allow users to create accounts and profiles. Examples include `webmd.com`, `mayoclinic.org`, `medscape.com`, `drugs.com`, `psychologytoday.com`. Therefore, the third-parties for these website may have access to the detailed health-related information of users. The Electronic Frontier Foundation provides some evidence for this, where a health-related website shared users' health related data with third-parties (Quintin 2015). This makes the third-party sharing behavior of the health-related website much more sensitive. Moreover, even though there are other websites that do not necessarily include user-specific data, user behavior on such websites can still be sensitive, as it can be used to predict health-related facts about a user. For example, consider a user looking at a heart condition page on `cdc.gov` or `heart.org`. Such a user can be predicted to be of high risk for heart conditions by third-party tracker. Use of third-party cookies enables trackers to collect such information and depict a profile of the user with the information on their behavior in other websites. Therefore, compared to news websites, health-related websites are more sensitive, irrespective of whether they contain user-specific data or not. We now include this clarification in the paper in the following quote:

> "Alexa Internet provides rankings for publisher websites within different website subject categories. From these categories, we take the list of the 500 most-visited websites from two categories: News and Health. These two categories were selected with the intention of finding website subject categories for which users have different privacy concerns and intentions to disclose personal information and browsing behavior due to the nature of the subject content. Health information is generally thought to have higher content sensitivity than News. Many websites allow users to create accounts and profiles. This is particularly important for Health websites, for example `webmd.com`, `mayoclinic.org`, `medscape.com`, `drugs.com`, `psychologytoday.com`. The third-parties used by these website have access to the detailed health-related information of users. The Electronic Frontier Foundation provides some evidence for this, where a Health website shared users' health related data with third-parties (Quintin 2015). This makes the third-party sharing behavior of Health website much more sensitive than News websites. Even though not all websites collect user-specific data, user behavior on Health websites can still be sensitive, as it can be used to predict health-related facts about a user. For example, consider a user looking at a heart condition page on `cdc.gov` or `heart.org`. Such a user can be predicted to be of high risk for heart conditions by third-party trackers. Use of third-party cookies enables trackers to collect such information and depict a profile of the user with the information on their behavior in other websites. Therefore, compared to News websites, Health websites are more sensitive in terms of privacy concern. This line of reasoning is consistent with previous studies on website categories and content sensitivity, including Gopal et al. (2018). "

Overall, I like the current draft of the paper.

We appreciate your feedback as well as your support, and are happy to see that you were content with the previous revision of the paper. Your latest comments have helped to further improve the paper.

Reviewer 2's Comments

First of all, I would like to compliment the authors on the substantial effort invested in this revision. The theory part in the current version now presents a completely overhauled static incomplete information game and its equilibrium solution. While simplifications have been made to make this possible within the tight time frame, I agree that this is the right strategy to take. In comparison to the first version, the level of information sharing, m, is now exogenous. The scope of strategic decision making is solely on obfuscation, η. The theoretical model is now much more simplistic and tractable, and yet still offers what is needed to guide the empirical analysis. This is no doubt a success.

Thank you for your kind words. Your deep insights on developing a more parsimonious model that is consistent with the empirical analysis were very helpful in revising the paper.

While I really like models that are simple and elegant, I wish to point out that in reality, the level of information sharing is more likely determined at the same time as the level of obfuscation. Websites by dynamically deciding on which third parties to sell user information to at each refresh/click, effectively determines both the level of obfuscation and the level of information selling. Theses two are really intertwined. Moreover, users often have an idea about the level of "prominence" or "value" and the level of "content sensitivity" of a website that they intend to visit. There is scope in improving the modelling to a) incorporate these features and b) reflect the nature that websites of contrasting values and content sensitivities receive traffic from users of different characteristics. That is, users often "observe" website characteristics before their decision on usage and counter obfuscation effort. More privacy concerned users may be more likely to shy away from websites that engage in more intensive monetization through sharing user information with third parties, based on whether they are vising a prominent (famous vs less known) or a content sensitive (news stories or health related articles) website. This warrants a dynamic model of incomplete information.

While a simple static model is preferred as long as essential insights are preserved, we agree with you that dynamic incomplete information model can provide additional insights and realism. We believe this to be a good avenue for future research and have included the following comment in the revised paper.

> "We provide a simple and elegant model to capture the nuances involved in the obfuscation of third-parties. That said, in reality, the level of information sharing is more likely determined at the same time as the level of obfuscation in a dynamic manner. This makes a dynamic model of incomplete information a useful avenue for future research on this topic."

Regarding the now also improved empirical analysis, could I asked about "obfuscation" in the absence of DNT? My understanding is that even without DNT being selected in the browser, some websites also add and/or drops third parties upon each refresh. Moreover, upon a new refresh/click the website may find it more valuable to sell user information to "new" third parties instead of those already sold to on the user's previous visits. Indeed, this simple logic would mean that adding and dropping websites upon each request to the server is a practice that has something to do with the extent of information selling rather than obfuscation.

This brought me thinking 1) why the paper needed to record websites response to DNT to study obfuscation. Isn't obfuscation also relevant without DNT? It seems plausible that some users who value privacy do not know there's a DNT option in their browser or, in contrast, some users won't be bothered to enable DNT as they know websites are not legally obliged to respect it.

Your observation is correct and very insightful. It is true that websites could change their set of third-parties with each visit. This is the reason that in our empirical work we visit each website multiple times to create a more cohesive picture of the third-parties used under each browsing option (with and without DNT). Obfuscation is relevant to users both with and without DNT. The use of DNT is a way of detecting differential obfuscation based on the user action of enabling DNT. Your point about why some users who value privacy do not choose the DNT option is well taken. In our experimental set up, we programmatically set the DNT option in order to expose the website response to user action.

And 2) how do we know obfuscation is not a different form of "abuse"? After all, there is not much harm to profits in dropping third parties with whom the website already shared information of this user.

This is indeed the case, and our definition of obfuscation is based on this as well. Obfuscation is defined as "practices and strategies that firms employ to confuse users and prevent them from recognizing the best offer (Ellison and Wolitzky 2012 and Gu and Wenzel 2014)". The difference between abuse and obfuscation is that abuse is where the website disregards the users' DNT request and tracks them even more, whereas obfuscation is the change in response to DNT in third-parties in a way that the reaction to DNT is not clear. In other words, obfuscation makes it unclear so as to what the website's reaction to DNT is. We note that these definitions are consistent with Merriam-Webster's definitions of obfuscation (to be evasive, unclear, or confusing) and abuse (a corrupt practice or custom).

To be clear, I am not demanding any changes to the draft as a result of these comments. I am happy to see this paper published as it is. I can see clear value in both the theoretical and empirical analyses of the paper as a first study of its kind. Thanks again for taking my comments on the previous version on board.

Thank you for all of your very insightful and helpful comments in both this round and the previous round. The comments provided by yourself, as well as the other reviewers, have indeed helped us a great deal to substantially improve the paper.

P.S., in case it is not picked up by others, in line 24 on page 27, it should have been "Further" instead of "further".

Thank you. This typo has been fixed. We have also done a thorough proof-reading of the new revision.

Reviewer 3's Comments

In this revised manuscript, the authors rebuild a Bayesian model and consider more user heterogeneity aspects to help understand the obfuscation of information sharing in websites with different content sensitivity types. The relationship and concept consistency between the analytical model and corresponding empirical analyses are strengthened. I thank the authors for the efforts in improving the manuscript. Some concerns and comments towards the revisions are as below and I hope the authors find them useful.

Thank you for all of your comments in both this round and the previous round. The comments provided by yourself, as well as the other reviewers, have indeed helped us a great deal to substantially improve the paper.

1. For the point 1 in the last round review towards the literature review, some key prior literatures could be better summarized and reviewed with more details, especially for the ones related to the obfuscations of information sharing and effects of third parties on information diffusions. Similarly, some cited papers, such as the ones focusing on website linkages through user preferences and user online shopping patterns, should be better elaborated for their relations to the main topic of this research.

Thank you for your comment. In this revision, we have overhauled the literature review section. We have included and discussed related work on obfuscation in the context of firms and product search. It is important to note that the obfuscation of information sharing by websites is a novel problem that we explore in this paper, and therefore, we were not able to find any papers that are directly addressing this issue. Additionally, as you suggested, we have expanded on the relationship between our work and prior work on website third-parties, and have elaborated the contribution of our work to each literature stream in more detail. The new literature review section is copied here for your convenience:

> "The technical implications of third-parties and their impact on users' information diffusion and leakage have been studied using a variety of methods such as crowd-sourcing (Yu et al. 2016) and web crawling (Englehardt and Narayanan 2016). Krishnamurthy et al. (2011) study the websites with user registration, and find 75% of the websites to leak sensitive user information to third-parties. Roesner et al. (2012) detect and classify five different types of third-party trackers based on how they work within the browser environment. Acar et al. (2020) study the extent of data collection by third-parties, investigating the scripts that are directly embedded on web pages. We extend this literature by studying the website's obfuscation of third-party usage among websites. Our analysis shows that obfuscation is popular among popular websites.
>
> Users' browsing and personal data is shared with third-parties by websites, mainly for the purpose of monetization and advertising. This is done through use of third-party cookies and passing data through *https* requests (Libert 2015, Englehardt et al. 2015, Englehardt and Narayanan 2016). There have been several studies on the relationship between websites and third-parties, and the underlying mechanisms for websites to improve their usability (Gupta et al. 2007) and monetization (Gopal et al. 2018) through third-party trackers. Gopal et al. (2018) study the website's trade-off between sharing user data with third-parties and user privacy concerns. Gupta et al. (2007) propose a methodology to improve the linkage of websites according to user preferences estimated based on user data, and Yang et al. (2013) offer a framework to discover users online shopping patterns across websites using their online behavior. The interaction of websites and third-parties forms an integral part of the website's monetization of users, which we study in this paper. Rather than focusing on the operational details of websites in terms of their linkage and discovery of users, we focus on the incentive of websites to obfuscate their use of third-parties.

> Use of third-parties comes with implications for privacy. Online privacy has been a subject of many prior studies. Smith et al. (2011), Pavlou (2011), and Li (2012) provide a comprehensive review of the extensive online information privacy literature and develop frameworks for theoretical research on user privacy decision making. Bai et al. (2012) consider the security and privacy issues at organizational workflows and suggest consideration to improve exposure. We directly include privacy in our analysis, and extend it to include the interaction between privacy concern of users, content sensitivity of websites, and the level of information sharing at the website. This allows us to capture the different factors that drive the total privacy cost that users face.
>
> Information asymmetry surrounds our analysis of obfuscation. Given that users do not readily know the extent of information sharing, there is information asymmetry between website and users. Websites can utilize this to their benefit. Such implications of information asymmetry have been extensively studied, for example in the context of quality and uncertainty (Akerlof 1970), moral hazard (Hölmstrom 1979), and agency (Jenson and Meckling 1976, Arrow 1985, Eisenhardt 1989). We extend this by including the ability of websites to change the level of information asymmetry, which is done through obfuscation.
>
> Obfuscation encompasses strategies to increase the information asymmetry between the website and users by making it costly to discover the true extent of website information sharing with third-parties. Researchers have studied obfuscation as a way for firms to prevent customers from recognizing the best offer (Ellison and Wolitzky 2012, Gu and Wenzel 2014), and this is consistent with our view of obfuscation by websites. Obfuscation of information has also been studied in the context of firms' information sharing with stakeholders within firm disclosures (Bushee et al. 2018), higher education signaling (Goldstein and Eaton 2021), and product and price search results (Wilson 2010). However, to the best of our knowledge, our analysis is the first to consider the obfuscation carried out by websites to make their sharing of user information less transparent."

2. Related to the point 8 in the last round review, user information could be shared to all linked third parties, no matter it's a primary one or secondary one. Thus, only using the primary ones to indicate the true information sharing level could be biased. Incorporating the secondary third parties can better reflect the true information sharing level and address the bias concern to a certain extent. However, the robustness checks do not present very solid and consistent supports to the main findings. The explanation for the insignificant effects of Rank is reasonable and acceptable. While some further analyses and discussions towards the positive effects of Category on Abuse in Column 1 of Table 8 should be carefully conducted and analyzed, as this is one key conclusion in this paper. Statistics and illustrations for the dynamic changes of secondary third parties through 10 runs can be presented for a better understanding of the fluctuations and distributions of uncontrolled secondary third parties.

As you suggested, we have thoroughly reanalyzed our models and the robustness of these models. The robustness results are now presented in Table 8 for the total number of third parties. Comparing Table 7 (main results using primary third parties) and Table 8 shown below demonstrate that the results are consistent.

Table 7. Website Reaction to DNT (Two-stage Model) Using Primary Third-Parties

	First Stage		Second Stage	
	$Abuse$	$Abuse - Level$	$Obfuscate$	$Obfuscation - Level$
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
$Category$	-1.542***	-1.530***	-2.656*	-0.585***
	(0.290)	(0.272)	(1.241)	(0.217)
$Rank$	0.000	-0.001	-0.007***	-0.001***
	(0.001)	(0.001)	(0.002)	(0.000)
Ad_{res}	0.453***	0.421***	-0.987	0.141
	(0.144)	(0.135)	(0.673)	(0.096)
$Login$	0.398	0.500	0.048	0.091
	(0.336)	(0.316)	(0.807)	(0.217)
$Prim_{res}$	0.327	-0.100	1.336	-0.241
	(0.344)	(0.319)	(1.177)	(0.250)
Constant	-0.736*	-0.756**	0.195	0.946
	(0.394)	(0.365)	(0.979)	(0.231)
Observations	400	400	93	93
AIC	393.349	570.169	94.972	294.481
Hosmer-Lemeshow p-value	0.599	0	0.439	1.000
McFadden's R^2	0.121	0.082	0.184	0.051
Cox & Snell R^2	0.123	0.117	0.183	0.151
Cameron & Windmeijer R^2	0.121	0.191	0.184	0.225

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table 8. Website Reaction to DNT (Two-stage Model) Using Total (Primary+Secondary) Third-Parties

	First Stage		Second Stage	
	$Abuse$	$Abuse - Level$	$Obfuscate$	$Obfuscation - Level$
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
$Category$	-0.674***	-1.512***	-2.591***	-1.305***
	(0.246)	(0.275)	(0.599)	(0.145)
$Rank$	0.001	-0.001	-0.001	-0.001**
	(0.001)	(0.001)	(0.002)	(0.0005)
Ad_{res}	0.504***	0.563***	1.111***	0.452***
	(0.128)	(0.133)	(0.299)	(0.074)
$Login$	0.282	0.451	0.116	0.116
	(0.287)	(0.321)	(0.614)	(0.165)
$Total_{res}$	-0.024	0.586*	3.597***	1.149***
	(0.288)	(0.309)	(0.857)	(0.158)
Constant	-1.079***	0.761*	2.692***	3.007***
	(0.354)	(0.392)	(0.841)	(0.202)
Observations	400	400	93	93
AIC	462.831	1045.172	102.489	747.015
Hosmer-Lemeshow p-value	0.069	1.000	0.016	1.000
McFadden's R^2	0.064	0.042	0.377	0.113
Cox & Snell R^2	0.074	0.108	0.376	0.552
Cameron & Windmeijer R^2	0.064	0.171	0.377	0.537

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

3. For my points 13 and 14 towards the sample selection in the last round review, it's not convincible to remove websites with unknown average number of monthly users, as this variable is not further used in the empirical analysis. Or this variable is for the Rank proxied by the information from Alexa? If no, this variable may be employed to reflect Rank as a robustness check. Also, in my opinion, I don't think it's necessary to randomly select 200 websites in each category to keep the balanced list, and it's adaptable to use the total of 676 websites (391 News and 285 Health) for empirical analyses.

To make data collection more manageable we have decided to limit our analysis to 400 websites. We narrowed our list by selecting a random sample of 200 websites with fully available information from each category. We believe that 200 websites in each category is representative of each category, and therefore, did not collect data on the rest of the websites. Given our large sample size of 200 for each category, we have no reason to expect that a larger sample size would impact our results.

We used website *Rank* as a proxy for website value, because we believe it captures the quality of website better than the average number of users (*Size*). As you suggested we also employed the average number of monthly users (*Size*) to reflect *Rank* as a robustness check. The results are consistent with that of the main model. We now include this robustness check in the paper as follow:

> "We also employed the average number of monthly users (*Size*) to reflect *Rank* as a robustness check. We replaced *Rank* with $log(Size)$ in all models. Results are provided in Table 9 (for primary third-parties) and Table 10 (for total third-parties). It can be seen that both tables are consistent with the main results."

Table 9. Website Reaction to DNT (Two-stage Model With log(Size)) Using Primary Third-Parties

	First Stage		Second Stage	
	Abuse	*Abuse − Level*	*Obfuscate*	*Obfuscation − Level*
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
Category	-1.442***	-1.387***	-2.178*	-0.453**
	(0.291)	(0.272)	(1.233)	(0.215)
$log(Size)$	0.190**	0.209***	0.589***	0.144***
	(0.088)	(0.075)	(0.185)	(0.046)
Ad_{res}	0.442***	0.407***	0.949	0.129
	(0.144)	(0.134)	(0.677)	(0.097)
Login	0.381	0.459	-0.166	0.039
	(0.339)	(0.314)	(0.834)	(0.216)
$Prim_{res}$	0.289	-0.023	1.518	-0.159
	(0.349)	(0.318)	(0.977)	(0.252)
Constant	-3.457***	-3.836***	-9.064***	-1.309*
	(1.237)	(1.073)	(2.747)	(0.681)
Observations	400	400	93	93
AIC	389.616	564.693	94.535	292.251
Hosmer-Lemeshow p-value	0.494	1.000	0.652	1.000
McFadden's R^2	0.129	0.092	0.189	0.059
Cox & Snell R^2	0.131	0.129	0.187	0.171
Cameron & Windmeijer R^2	0.129	0.210	0.188	0.259

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table 10. Website Reaction to DNT (Two-stage Model With log(Size)) Using Total Third-Parties

	First Stage		Second Stage	
	$Abuse$	$Abuse-Level$	$Obfuscate$	$Obfuscation-Level$
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
$Category$	-0.715***	-1.452***	-2.506***	-1.252***
	(0.248)	(0.277)	(0.594)	(0.143)
$log(Size)$	-0.040	0.128	0.193	0.118***
	(0.084)	(0.094)	(0.204)	(0.044)
Ad_{res}	0.505***	0.560***	1.110***	0.448***
	(0.127)	(0.132)	(0.296)	(0.074)
$Login$	0.308	0.351	-0.034	0.038
	(0.286)	(0.318)	(0.652)	(0.168)
$Total_{res}$	-0.017	0.577*	3.490***	1.147^***
	(0.286)	(0.309)	(0.847)	(0.155)
Constant	-0.327	-1.035	0.054	1.276**
	(1.161)	(1.307)	(2.612)	(0.580)
Observations	400	400	93	93
AIC	463.924	102.555	1043.890	745.036
Hosmer-Lemeshow p-value	0.675	0.020	1.000	1.000
McFadden's R^2	0.062	0.376	0.044	0.115
Cox & Snell R^2	0.072	0.376	0.111	0.560
Cameron & Windmeijer R^2	0.062	0.376	0.176	0.544

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

4. Related to my review point 7 towards the multicollinearity concern in the last round, it's interesting to see the empirical results for estimating coefficients in Equations (8) and (9) directly, considering that the VIF tests results are acceptable.

As you suggested, we analyze the empirical results for estimating coefficients in Equations (8) and (9) directly. Results of the one-stage models have been provided in Table 11. The effect of $Category$ on $Abuse$ and $Obfuscate$ are significant, consistent with the main results. However the effect of $Category$ on $Abuse-level$ and $Obfuscation-level$ are not significant. The effect of $Rank$ on $Obfuscation-Level$ is significant, consistent with the main results. However the effect of $Rank$ on $Obfuscate$ is not significant. Overall, the one-stage model results are fairly consistent with the analytical model findings in Proposition 1 and Proposition 2. We added these results to the "Robustness Check" section in the paper as follows:

> "Moreover, we analyze the empirical results for estimating coefficients in Equations (8) and (9) directly. Results of the one-stage models are provided in Table 11. The effect of $Category$ on $Abuse$ and $Obfuscate$ are significant, consistent with the main results. However the effect of $Category$ on abuse and obfuscation levels are not significant. The effect of $Rank$ on $Obfuscation-Level$ is significant, consistent with the main results. However the effect of $Rank$ on $Obfuscate$ is not significant. While some of the one-stage model results are different from two-stage models, the results for the main coefficients are consistent with the analytical model findings in Proposition 1 and Proposition 2."

Table 11. Website Reaction to DNT (One-stage Model) Using Primary Third-Parties

	First Stage		Second Stage	
	$Abuse$	$Abuse - Level$	$Obfuscate$	$Obfuscation - Level$
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
Category	-0.754**	-1.121	-0.744**	-0.279
	(0.346)	(0.891)	(0.322)	(0.232)
Rank	-0.001	-0.007***	-0.001	-0.002***
	(0.001)	(0.002)	(0.001)	(0.001)
Advertising	1.300***	17.353	1.305***	0.519*
	(0.386)	(1587.729)	(0.369)	(0.294)
Login	0.269	-0.116	0.325	0.033
	(0.337)	(0.741)	(0.315)	(0.211)
Constant	-1.866***	-16.860	-1.782***	0.552
	(0.504)	(1587.730)	(0.473)	(0.368)
Observations	400	400	93	93
AIC	391.651	566.100	92.946	292.385
Hosmer-Lemeshow p-value	0.537	0	0.718	1.000
McFadden's R^2	0.120	0.086	0.184	0.052
Cox & Snell R^2	0.122	0.122	0.183	0.152
Cameron & Windmeijer R^2	0.120	0.198	0.185	0.227

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

5. The goodness-of-fit values, such as R square or pseudo R square values, for empirical analyses should be reported in Tables 7 and 8. Related statistics for the variable Primary should be presented in Tables 4 and 5.

Thank you for your suggestion on this. Related statistics for the Variable *Primary* have been added to Tables 4 and 5, and several goodness-of-fit values and pseudo R squares have been added to all tables.

6. How to calculate the obfuscation level in the empirical analysis? Is it the number of newly added third parties or the number of changed third parties covering both adding and dropping ones?

Thank you for this question. Our definition of the obfuscation level was indeed inadequate in the previous revision. $Obfuscation - Level$ is the number of changed third-parties which includes those that are both "added with DNT" and "dropped with DNT". For example if the website adds 3 and drops 2, the total change is counted as 5. Note that this is only calculated for the websites that choose to abuse in the first stage, which means that the number of added third-parties is higher than the number of dropped ones. To address this confusion, the following paragraph has been added to the paper to clarify the definitions of $Obfuscation - Level$ and $Abuse - Level$.

> "We also examine the magnitude of websites' reaction to user privacy request. We capture this using two additional dependent variables, $Abuse - Level$ and $Obfuscation - Level$. $Abuse - Level$ is the number of third-parties added in response to user privacy request, and $Obfuscation - Level$ is the summation of both the number of third-parties added and dropped with the user privacy request. Because these additional dependent variables are count values, we analyze them using negative binomial regression model."

7. Some typos should be carefully checked, such as the inconsistent 0/1 definitions for Abuse in Table 3 and empirical analyses and some citation formats in the main text.

Thank you for noting these typos and issues. We have addressed these in this revision. Additionally, we have done a thorough proof-reading of the paper to remove any other typos and grammatical errors.

References

Acar G, Englehardt S, Narayanan A (2020) No boundaries: data exfiltration by third parties embedded on web pages. *Proceedings on Privacy Enhancing Technologies* 2020(4):220–238.

Akerlof G (1970) The market for lemons: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics* 84(3):488–500.

Arrow KJ (1985) The economics of agency. Technical report, In J. W. Pratt R. J. Zeckhauser (Eds.), Principals and agents: The structure of business (pp. 37-51). Boston: Harvard Business School Press.

Bai X, Gopal R, Nunez M, Zhdanov D (2012) On the prevention of fraud and privacy exposure in process information flow. *INFORMS Journal on Computing* 24(3):416–432.

Bushee BJ, Gow ID, Taylor DJ (2018) Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research* 56(1):85–121.

Eisenhardt KM (1989) Agency theory: An assessment and review. *Academy of management review* 14(1):57–74.

Ellison G, Wolitzky A (2012) A search cost model of obfuscation. *The RAND Journal of Economics* 43(3):417–441.

Englehardt S, Narayanan A (2016) Online tracking: A 1-million-site measurement and analysis. *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 1388–1401.

Englehardt S, Reisman D, Eubank C, Zimmerman P, Mayer J, Narayanan A, Felten EW (2015) Cookies that give you away: The surveillance implications of web tracking. *Proceedings of the 24th International Conference on World Wide Web*, 289–299.

Goldstein A, Eaton C (2021) Asymmetry by design? identity obfuscation, reputational pressure, and consumer predation in us for-profit higher education. *American Sociological Review* 86(5):896–933.

Gopal RD, Hidaji H, Patterson RA, Rolland E, Zhdanov D (2018) How much to share with third parties? user privacy concerns and website dilemmas. *MIS Quarterly* 42(1):143–164.

Gu Y, Wenzel T (2014) Strategic obfuscation and consumer protection policy. *The Journal of Industrial Economics* 62(4):632–660.

Gupta R, Bagchi A, Sarkar S (2007) Improving linkage of web pages. *INFORMS Journal on Computing* 19(1):127–136.

Hölmstrom B (1979) Moral hazard and observability. *The Bell journal of economics* 74–91.

Jenson MC, Meckling WH (1976) Theory of the firm: managerial behavior, agency costs and ownership structure. *Journal of financial economics* 3(4):305–360.

Krishnamurthy B, Naryshkin K, Wills C (2011) Privacy leakage vs. protection measures: the growing disconnect. *Proceedings of the Web*, volume 2, 1–10.

Li Y (2012) Theories in online information privacy research: A critical review and an integrated framework. *Decision Support Systems* 54(1):471–481.

Libert T (2015) Exposing the hidden web: An analysis of third-party http requests on 1 million websites. *arXiv preprint arXiv:1511.00619* .

Pavlou PA (2011) State of the information privacy literature: Where are we now and where should we go? *MIS Quarterly* 977–988.

Quintin C (2015) Healthcare. gov sends personal data to dozens of tracking websites. *Electronic Frontier Foundation* .

Roesner F, Kohno T, Wetherall D (2012) Detecting and defending against third-party tracking on the web. *Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*, 155–168.

Smith HJ, Dinev T, Xu H (2011) Information privacy research: an interdisciplinary review. *MIS Quarterly* 35(4):989–1016.

Wilson CM (2010) Ordered search and equilibrium obfuscation. *International Journal of Industrial Organization* 28(5):496–506.

Yang Y, Liu H, Cai Y (2013) Discovery of online shopping patterns across websites. *INFORMS Journal on Computing* 25(1):161–176.

Yu Z, Macbeth S, Modi K, Pujol JM (2016) Tracking the trackers. *Proceedings of the 25th International Conference on World Wide Web*, 121–132.

INFORMS Journal on Computing

Now You See it, Now You Don't: Obfuscation of Online Third-Party Information Sharing

Ashkan Eshghi
Haskayne School of Business, University of Calgary, Calgary, Alberta T2N 1N4, Canada, ashkan.eshghi1@ucalgary.ca

Ram Gopal
Warwick Business School, University of Warwick, ram.gopal@wbs.ac.uk

Hooman Hidaji, Raymond Patterson
Haskayne School of Business, University of Calgary, Calgary, Alberta T2N 1N4, Canada,
hooman.hidaji@haskayne.ucalgary.ca, raymond.patterson@ucalgary.ca

The practice of sharing online user information with external third-parties has become the focal point of privacy concerns for consumer advocacy groups and policy makers. We explore the decisions by websites regarding the obfuscation that they employ to make it difficult for users to discover the extent of information sharing. Using a Bayesian model, we shed light on the websites' incentive to obfuscate user information sharing. We find that as content sensitivity increases, a website reduces its level of obfuscation. Further, more popular websites engage in higher levels of obfuscation than less popular ones. We provide an empirical analysis of obfuscation and user information sharing in News (low content sensitivity) and Health (high content sensitivity) websites and confirm key results from our analytical model. Our analysis illustrates that obfuscation of information sharing is a viable strategy that websites use to improve their profits.

Keywords: third-parties, information sharing, obfuscation, privacy, information asymmetry

> *"Unregulated free markets rarely reward the different kind of heroism, of those who restrain themselves from taking advantage of customers' psychological or informational weaknesses."*
>
> Akerlof and Shiller (2016)

1. Introduction

Websites are the crux of the Internet and online activity. A vast array of services and content provided by the websites are made possible through third-parties. Third-parties provide a wide range of services and resources to websites, ranging from basic functionality such as hosting content to monetization capabilities such as advertising. The website shares user information with third-parties when rendering the service, which enables it to monetize the users. This can include basic browsing data such as IP addresses and browser and device information, as well as personal information such as marital status, income, and interests. As users have become aware of the implications of the lack of online privacy, third-parties have become an important point of argument among the public, policy makers, and businesses. It is therefore crucial to understand the implications of the use of third-parties for privacy, and the mechanisms that affect the use of third-parties as well as the transparency of their use.

There is information asymmetry between users and websites in the sense that the website's user information sharing is not always transparent, and thus users cannot readily determine the level of information sharing with third-parties. This information asymmetry between website and users creates an opportunity for the website to take advantage of users' informational weaknesses and utilize strategies that favor their own interest (refer to Akerlof (1970), Hölmstrom (1979), Jenson and Meckling (1976), Arrow (1985), and Eisenhardt (1989)). While the extent of information sharing is websites' private information, users can expend some effort and monitor websites' sharing behavior to reduce or remove the information asymmetry. Depending on the benefits from information asymmetry, which are mediated by content sensitivity and user privacy concern, it may be profitable for the website to hide its information sharing behavior to mislead users and increase the cost of monitoring. We refer to this as *obfuscation.* According to Akerlof and Shiller (2015), uninformed online users, or “Information Phools” make decisions based on information that is intentionally crafted to mislead them.

Focusing on websites' obfuscation of user information sharing with third-parties, we address the following research questions: 1) What are the website's incentives to obfuscate their sharing of user information with third-parties? 2) What type of websites are more prone to obfuscating their sharing of user information? To answer these questions, we first develop an analytical model based on the Bayesian equilibrium framework to capture and analyze the information asymmetry problem in the relationship between users and websites. We consider a monopolist website that is free to users, but charges third-parties for access to user information, and determines the level of obfuscation that it deploys. Our model sheds light on website's incentives to obfuscate ~~of~~ user information sharing. Interestingly, we find that website's obfuscation decreases as content sensitivity increases. Additionally, obfuscation increases as the website's value to users increases, implying that prominent websites are more prone to obfuscation. Further, we verify the key findings of our model through an empirical analysis of data from 400 News and Health websites. Our empirical analysis, consistent with our model, shows that content sensitive websites (Health) do not obfuscate as much as less sensitive websites (News), and that prominent websites obfuscate more than non-prominent ones.

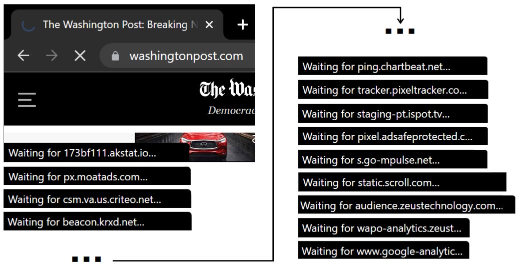
This paper contributes to our understanding of the websites' incentives, including those that drive websites to obfuscate their information sharing with third-parties, which has important implications for the information asymmetry between websites and users (Akerlof and Shiller 2015). Data regulation and privacy policy-makers can use our findings to design better policies which consider the websites' incentives to reduce the transparency of user information sharing. By clarifying the driving forces of website obfuscation, we contribute to the streams of literature on website monetization strategies (Gopal et al. 2018), online privacy (Li 2012, Pavlou 2011), obfuscation (Ellison and Wolitzky 2012, Gu and Wenzel 2014), and information flow and diffusion (Bai et al. 2012). ~~, and identification of users online (Yang et al. 2013).~~

It is useful to provide some context for the information asymmetry problem between a website and its users, including privacy concerns, sharing of user information with third-parties, and its obfuscation by websites. Users have been shown to take action to alleviate their online privacy concerns. According to Cisco (2021), in 2021, 86% of online users care about their data privacy and 79% of those who care are willing to act, with 47% of these users actually acting to preserve their data privacy. Thus, 32% of users took active steps to maintain their online privacy, increasing by 3% from 2020 (Cisco 2021). Concerns over privacy are found to have caused 48% of respondents from "a survey of over 8,000 consumers and business buyers across 16 countries" to either stop buying from a company or using their service (Salesforce 2019). In the context of websites, users can carefully observe visible privacy signals (e.g., privacy policies and safety badges). However, a survey conducted in 2019, shows that only 8% of US adults understand privacy policies (PewResearch 2019). A research by Cisco (2021) shows that 36% of online users lack trust that companies are truthfully following their stated privacy policies.

Users can utilize monitoring tools to understand the level of information sharing and make decisions that protect their online privacy. There are privacy tools and services that give some indication of the abusive behavior of websites. These privacy tools either depend on a community of users to monitor and tag abusive behavior (e.g., Web of Trust, Webutation, and Avast Web Reputation Plugin) or keep track of the third-parties that are engaged by the website (e.g., Blacklight, Lightbeam, Privacy Badger, and Ghostery). As an indication of widespread use of privacy tools, one of these tools, Web of Trust, has more than 140 million users (TechTimes 2020). Some examples of privacy tools and services are presented in Table 1.

Table 1 Examples of Using Privacy Tools for an Example Website (`washingtonpost.com`)

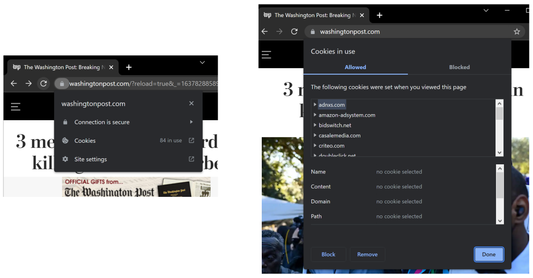
Browser loading notifications



The Washington Post: Breaking N
washingtonpost.com

Waiting for 173bf111.akstat.io...
Waiting for px.moatads.com...
Waiting for csm.va.us.criteo.net...
Waiting for beacon.krxd.net...
...
...
Waiting for ping.chartbeat.net...
Waiting for tracker.pixeltracker.co...
Waiting for staging-pt.ispot.tv...
Waiting for pixel.adsafeprotected.c...
Waiting for s.go-mpulse.net...
Waiting for static.scroll.com...
Waiting for audience.zeustechnology.com...
Waiting for wapo-analytics.zeust...
Waiting for www.google-analytic...

Google Chrome site information



The Washington Post: Breaking N
washingtonpost.com/?reload=true&_=16378288589

washingtonpost.com
Connection is secure
Cookies 84 in use
Site settings

The Washington Post: Breaking N
washingtonpost.com

Cookies in use

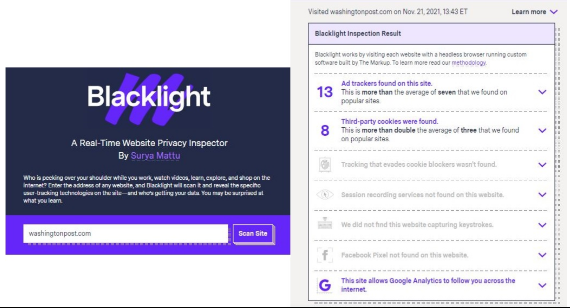
Allowed | Blocked

The following cookies were set when you viewed this page

- adnxs.com
- amazon-adsystem.com
- bidswitch.net
- casalemedia.com
- criteo.com
- doubleclick.net

Name no cookie selected
Content no cookie selected
Domain no cookie selected
Path no cookie selected

Blacklight



Blacklight

A Real-Time Website Privacy Inspector
By Surya Mattu

Who is peeking over your shoulder while you work, watch videos, learn, explore, and shop on the internet? Enter the address of any website, and Blacklight will scan it and reveal the specific user-tracking technologies on the site—and who's getting your data. You may be surprised at what you learn.

washingtonpost.com Scan Site

Visited washingtonpost.com on Nov. 21, 2021, 13:43 ET Learn more

Blacklight Inspection Result

Blacklight works by visiting each website with a headless browser running custom software built by The Markup. To learn more read our methodology.

13 Ad trackers found on this site. This is **more than** the average of **seven** that we found on popular sites.

8 Third-party cookies were found. This is **more than double** the average of **three** that we found on popular sites.

Tracking that evades cookie blockers wasn't found.

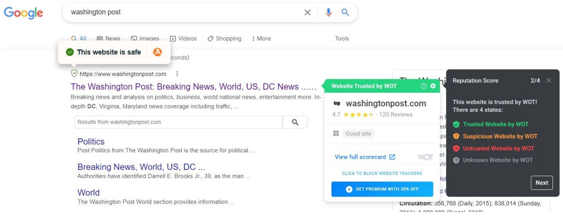
Session recording services not found on this website.

We did not find this website capturing keystrokes.

Facebook Pixel not found on this website.

This site allows Google Analytics to follow you across the internet.

Web of Trust – Avast



washington post

All News Images Videos Shopping More Tools

This website is safe

https://www.washingtonpost.com

The Washington Post: Breaking News, World, US, DC News ......
Breaking news and analysis on politics, business, world national news, entertainment more. In-depth DC, Virginia, Maryland news coverage including traffic, ...

Results from washingtonpost.com

Politics
Post Politics from The Washington Post is the source for political ...

Breaking News, World, US, DC ...
Authorities have identified Darrell E. Brooks Jr., 39, as the man ...

World
The Washington Post World section provides information ...

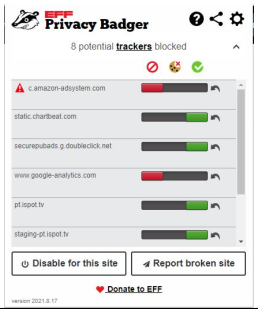
Website Trusted by WOT
washingtonpost.com
4.7 ★★★★☆ · 120 Reviews
Good site
View full scorecard
CLICK TO BLOCK WEBSITE TRACKERS
GET PREMIUM WITH 20% OFF

Reputation Score 2/4
This website is trusted by WOT! There are 4 states:
- Trusted Website by WOT
- Suspicious Website by WOT
- Untrusted Website by WOT
- Unknown Website by WOT

Next

Circulation: 356,768 (Daily, 2015); 838,014 (Sunday,

Privacy Badger



EFF Privacy Badger

8 potential **trackers** blocked

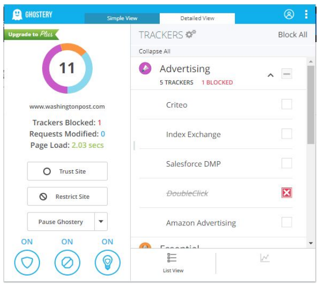
c.amazon-adsystem.com	🟥
static.chartbeat.com	🟩
securepubads.g.doubleclick.net	🟩
www.google-analytics.com	🟥
pt.ispot.tv	🟩
staging-pt.ispot.tv	🟩

Disable for this site | Report broken site

Donate to EFF

version 2021.8.17

Ghostery



GHOSTERY Simple View Detailed View

Upgrade to Plus

11

www.washingtonpost.com
Trackers Blocked: 1
Requests Modified: 0
Page Load: 2.03 secs

Trust Site
Restrict Site
Pause Ghostery

ON ON ON

TRACKERS Block All
Collapse All

Advertising
5 TRACKERS 1 BLOCKED

- Criteo ☐
- Index Exchange ☐
- Salesforce DMP ☐
- ~~DoubleClick~~ ☒
- Amazon Advertising ☐

Essential

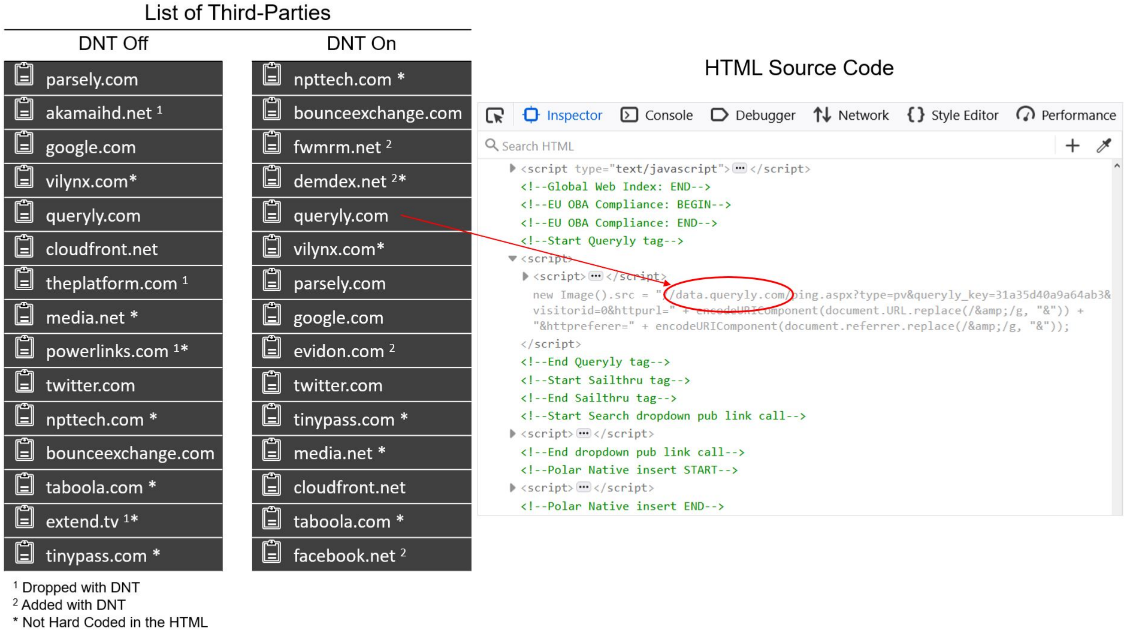
List View

On the other hand, websites can use obfuscation techniques to make monitoring more difficult. Websites use a variety of obfuscation techniques to impede users from understanding the true level of information sharing and its privacy implications. Obfuscation refers to practices and strategies that firms employ to confuse users and prevent them from recognizing the best offer (Ellison and Wolitzky 2012 and Gu and Wenzel 2014). One common technique employed by websites is use of "dark patterns", which tricks users into giving away their privacy (Vincent 2021). Such efforts by websites make it hard for users to discover the true level of information sharing and are instantiations of obfuscation strategies. In our context, obfuscation involves strategically preventing users from recognizing the true level of information sharing with third-parties.

For an example of website obfuscation of third-parties, consider the following. Users can utilize passive privacy tools, which signal to the website that such users are concerned about their privacy. One popular such tool is the passive request not to be tracked, known as a Do Not Track (DNT) request. Users may perceive the website's level of information sharing by comparing the number of third-parties used when the website is visited with and without a DNT request. On the other hand, websites can obfuscate the sharing of user information by making it hard for users to discern the reaction of websites to explicit user privacy requests. For example, some websites both add and drop third-parties that they share data with when users ask not to be tracked. This creates confusion about the true intent of the website and thus obfuscates the website's information sharing behavior. Moreover, websites can dynamically change the third-parties that are hard coded in the HTML source code. This creates even more confusion and makes it harder for privacy services to determine the true level of information sharing. To illustrate such obfuscation practices, we present the response of a website to a DNT request in Figure 1.

As shown in Figure 1, for this particular website, 15 third-parties are utilized. When the user enables DNT, the website drops four third-parties and replaces them with four new third-parties, leaving eleven third-parties unchanged. Adding to this state of confusion is the question of whether the third-parties can be discovered through the HTML source code of the website. In this illustration we find that 7 third-parties are not hard coded in the HTML source code when DNT is off, and 6 are not included in the HTML source code when DNT is on. This implies that if a user or an organization were to study the HTML for this website, they would not be able to discover all third-parties that are utilized. All

Figure 1 An Example of Website Information Sharing Obfuscation



List of Third-Parties

DNT Off	DNT On
parsely.com	npttech.com *
akamaihd.net [1]	bounceexchange.com
google.com	fwmrm.net [2]
vilynx.com*	demdex.net [2]*
queryly.com	queryly.com
cloudfront.net	vilynx.com*
theplatform.com [1]	parsely.com
media.net *	google.com
powerlinks.com [1]*	evidon.com [2]
twitter.com	twitter.com
npttech.com *	tinypass.com *
bounceexchange.com	media.net *
taboola.com *	cloudfront.net
extend.tv [1]*	taboola.com *
tinypass.com *	facebook.net [2]

[1] Dropped with DNT
[2] Added with DNT
* Not Hard Coded in the HTML

HTML Source Code

```
▸ <script type="text/javascript">⋯</script>
  <!--Global Web Index: END-->
  <!--EU OBA Compliance: BEGIN-->
  <!--EU OBA Compliance: END-->
  <!--Start Queryly tag-->
▾ <script>
  ▸ <script>⋯</script>
    new Image().src = "//data.queryly.com/ping.aspx?type=pv&queryly_key=31a35d40a9a64ab3&
    visitorid=0&httpurl=" + encodeURIComponent(document.URL.replace(/&amp;/g, "&")) +
    "&httpreferer=" + encodeURIComponent(document.referrer.replace(/&amp;/g, "&"));
  </script>
  <!--End Queryly tag-->
  <!--Start Sailthru tag-->
  <!--End Sailthru tag-->
  <!--Start Search dropdown pub link call-->
▸ <script>⋯</script>
  <!--End dropdown pub link call-->
  <!--Polar Native insert START-->
▸ <script>⋯</script>
  <!--Polar Native insert END-->
```

of this represents a great deal of change in reaction to the DNT request, which confuses users and privacy services about the true intention of the website.

2. Literature Review

The technical implications of third-parties and their impact on information diffusion and leakage have been studied using a variety of methods such as crowd-sourcing (Yu et al. 2016) and web crawling (Englehardt and Narayanan 2016). Krishnamurthy et al. (2011) study the websites with user registration, and find 75% of the websites to leak sensitive user information to third-parties. Roesner et al. (2012) detect and classify five different types of third-party trackers based on how they work within the browser environment. Acar et al. (2020) study the extent of data collection by third-parties, investigating the scripts that are directly embedded on web pages. We extend this literature by studying the website's obfuscation of third-party usage among websites. Our analysis shows that obfuscation is popular among popular websites.

Users' browsing and personal data is shared with third-parties by websites, mainly for the purpose of monetization and advertising. This is done through use of third-party cookies and passing data through *https* requests (Libert 2015, Englehardt et al. 2015, Englehardt and Narayanan 2016). There have been several studies on the relationship between websites and third-parties, and the underlying mechanisms for websites to improve their usability (Gupta et al. 2007) and monetization (Gopal et al. 2018) through third-party trackers. Gopal et al. (2018) study the website's trade-off between sharing user data with third-parties and user privacy concerns. Gupta et al. (2007) propose a methodology to improve the linkage of websites according to user preferences estimated based on user data, and Yang et al. (2013) offer a framework to discover users online shopping patterns across websites using their online behavior. The interaction of websites and third-parties forms an integral part of the website's monetization of users, which we study in this paper. Rather than focusing on the operational details of websites in terms of their linkage and discovery of users, we focus on the incentive of websites to obfuscate their use of third-parties.

Use of third-parties comes with implications for privacy. Online privacy has been a subject of many prior studies. Smith et al. (2011), Pavlou (2011), and Li (2012) provide a comprehensive review of the extensive online information privacy literature and develop frameworks for theoretical research on user privacy decision making. Bai et al. (2012) consider the security and privacy issues at organizational workflows and suggest consideration to improve exposure. We directly include privacy in our analysis, and extend it to include the interaction between privacy concern of users, content sensitivity of websites, and the level of information sharing at the website. This allows us to capture the different factors that drive the total privacy cost that users face.

Information asymmetry surrounds our analysis of obfuscation. Given that users do not readily know the extent of information sharing, there is information asymmetry between website and users. Websites can utilize this to their benefit. Such implications of information asymmetry have been extensively studied, for example in the context of quality and uncertainty (Akerlof 1970), moral hazard (Hölmstrom 1979), and agency (Jenson and Meckling 1976, Arrow 1985, Eisenhardt 1989). We extend this by including the ability of websites to change the level of information asymmetry, which is done through obfuscation.

Obfuscation encompasses strategies to increase the information asymmetry between the website and users by making it costly to discover the true extent of website information sharing with third-parties. Researchers have studied obfuscation as a way for firms to prevent customers from recognizing the best offer (Ellison and Wolitzky 2012, Gu and Wenzel 2014), and this is consistent with our view of obfuscation by websites. Obfuscation of information has also been studied in the context of firms' information sharing with stakeholders within firm disclosures (Bushee et al. 2018), higher education signaling (Goldstein and Eaton 2021), and product and price search results (Wilson 2010). However, to the best of our knowledge, our analysis is the first to consider the obfuscation carried out by websites to make their sharing of user information less transparent.

3. Analytical Model

In this section, we develop an analytical model to study the website-user relationship and analyze the website's decision regarding obfuscation of information sharing with third-parties. For convenience, a summary of our analytical model notations are shown in Table 2.

We consider a website that offers a free service to potential users, and charges third-parties for access to user information. We assume that the website is one of two types with respect to its level of information sharing, denoted as $t \in \{L, H\}$: low-type (L), or high-type (H). We assume there is a common belief that the website is low-type (L) with probability

Table 2 Variables

Notation	Definition
X	Value of the service that is provided by the website.
r	Individual IT illiteracy, which is defined as user's inability to discover website's true information sharing level, where $r \in [0,1]$.
s	Privacy sensitivity of website content (i.e., content sensitivity).
t	Website type with respect to its level of information sharing, where $t \in \{L, H\}$, $t = L$ is low, and $t = H$ is high.
v	Individual user's privacy concern, implying their concern for their information being shared with third-parties, where $v \in [\underline{v}, \overline{v}]$, $0 \leqslant \underline{v} < \overline{v}$.
m_t	Level of information sharing for website of type t, where $m_L < m_H$.
η_t	Obfuscation level of website of type t.
θ	Probability that website's type t is L, whereas probability that website's type t is H is $1 - \theta$.
$K_{v,r}$	Action of a user with privacy concern v and IT illiteracy r, where $K_{v,r} \in \{u, e, n\}$ and u denotes using the website without expending effort to discover the website, e denotes expending effort, and n denotes neither using the website nor expending effort.
$U^K(v, r, t)$	The utility from action K that a user with privacy concern v and IT illiteracy r gains from website type t.
$C(\eta_t)$	Obfuscation cost function.
$D_t(\eta_L, \eta_H)$	Demand that type t website obtains by choosing η_t.
Π_t	Profit of type t website.

θ and high-type (H) with probability $1 - \theta$. The website's level of information sharing is the website's private information, but users can expend effort to determine this.

We define obfuscation as the practice of the website to make it difficult for users to determine its type. If the website does not obfuscate and is fully transparent, the cost of effort to discover the website type (its level of information sharing) is zero. As the website increases obfuscation, it increases the cost of user discovery of the website type. The cost to the user of discovering the website type is a function of both the user's technical ability and the level of obfuscation employed by the website. The more ability a user has, the lower their cost of discovery, and vice-versa. We refer to a users' lack of technical ability as their *IT illiteracy*, denoted as r, which is assumed to be uniformly distributed $(r \in [0,1])$ across the population. The effort cost needed to discover the website type for a user with IT illiteracy r for website of type t is given as $r\eta_t$, where η_t is the obfuscation level of website of type t. Where $r = 0$, the cost of discovery is zero, corresponding to the lowest user IT illiteracy. Where $r = 1$, the effort needed to discover the website type is η_t, corresponding to the highest level of user IT illiteracy.

Utility that users gain from using the website depends on the value of the service, denoted as X, and the privacy cost of using the service. The privacy cost of using the website for users depends on the website's level of information sharing (website type, t), website content sensitivity (denoted as s), and users' privacy concern (denoted as v). The website's level of information sharing is denoted as m_t for a website of type t, where $m_L < m_H$. Therefore, the total privacy cost that a user with privacy concern v incurs if they use the website of type t with content sensitivity of s is given as svm_t.

Content sensitivity (s) relates to the website's content. It is believed that different types of websites have different levels of content sensitivity, where, for example, health-related information is more sensitive than news (Gopal et al. 2018). User privacy concern (v), on the other hand, is specific to a given user, and implies their concern for their information being shared with third-parties. Those with high privacy concern are more sensitive to their data being shared (Gopal et al. 2018). Users are heterogeneous with respect to their privacy concern (Chellappa and Shivendu 2007), and are assumed to have a privacy concern according to a uniform distribution $v \in [\underline{v}, \overline{v}]$. Note that our model accounts for two dimensions of user heterogeneity: IT illiteracy (r) and privacy concern (v). We assume these two dimensions to be independent of each other among users.[1]

We model the website-user relationship as an incomplete information game. The website chooses the level of obfuscation and simultaneously users decide to either use the website,

[1] This assumption does not qualitatively impact our result, as any correlation between user privacy concern and IT illiteracy can be superimposed on our findings to capture the results for any special case.

expend effort to discover the website type or not use the website. As discussed above, users choose from three possible choices:

1. Use the website without expending effort. We denote these users with u.

2. Expend effort to discover the website type (t). Given the website type, these users use the website if they receive positive utility from the website. We denote these users with e.

3. Do not use the website and do not expend effort. We denote these users with n.

Accordingly, the users' action set is $\{u, e, n\}$, and each individual user's action is denoted as $K_{v,r} \in \{u, e, n\}$. The utility that users gain from choosing each action for each website type is given as:

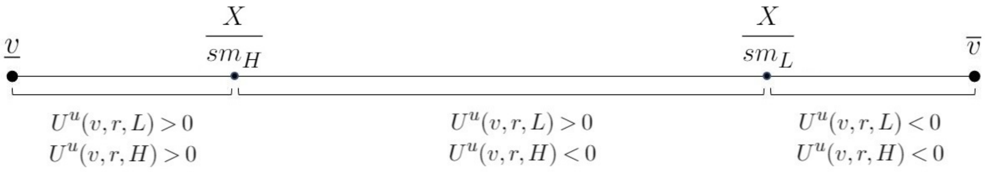
$$\begin{aligned}
&U^u(v,r,t) = X - svm_t \\
&U^e(v,r,t) = \begin{cases} X - svm_t - r\eta_t & \forall\, v \leqslant \frac{X}{sm_t} \\ -r\eta_t & \forall\, v > \frac{X}{sm_t} \end{cases} \qquad (1) \\
&U^n(v,r,t) = 0
\end{aligned}$$

Users with privacy concern $v \leqslant X/sm_H$ gain positive utility from using the website irrespective of the website's level of obfuscation and their IT illiteracy ($U^u(v,r,L) > 0$, $U^u(v,r,H) > 0$). Therefore, these users use the website without expending effort. On the other hand, users with $v \geqslant X/sm_L$ gain negative utility from using the website irrespective of the website's level of obfuscation and their IT illiteracy ($U^u(v,r,L) < 0$, $U^u(v,r,H) < 0$), and therefore, do not use the website and do not expend effort.[2] The analysis for users with $X/sm_H < v < X/sm_L$ (where $U^u(v,r,L) > 0$, $U^u(v,r,H) < 0$) is more complicated, as these users need to decide whether to use the website, not use the website, or expend

[2] For brevity, we focus on the more interesting scenario where $\underline{v} < X/sm_H < X/sm_L < \overline{v}$. Our results also apply to the case where this assumption does not hold.

effort to identify the website type, which then determines whether they use or not use the website. These decisions depend on the expected utility that these users receive from the website, as we explain next. Figure 2 provides an overview of the users' utility for users with different privacy concerns.

Figure 2 Users' Utility from Using the Website

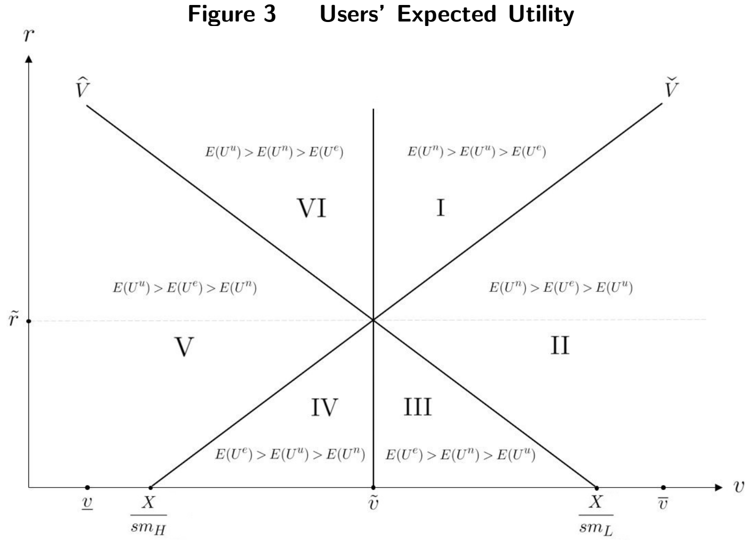


Based on (1), the expected utility of users (over the two possible website types) with $X/sm_H < v < X/sm_L$ is derived as a function of the expected level of information sharing, $E(m) = \theta m_L + [1-\theta]m_H$, and the expected level of obfuscation, $E(\eta) = \theta\eta_L + [1-\theta]\eta_H$. Using these definitions, we can derive the expected user utility as:

$$
\begin{aligned}
&E(U^u(v,r)) = \theta U^u(v,r,L) + [1-\theta]U^u(v,r,H) = X - svE(m)\\
&E(U^e(v,r)) = \theta U^e(v,r,L) + [1-\theta]U^e(v,r,H) = \theta[X - svm_L] - rE(\eta)\\
&E(U^n(v,r)) = 0
\end{aligned}
\tag{2}
$$

The expected user utility of use is simply the expected utility for the two website types. The expected utility of effort is composed of the weighted utility of using the website if it is low-type ($\theta[X - svm_L]$) subtracted by the expected cost of effort, noting that the utility of not using the website if it is high-type is zero. Based on the above utility expectations, we can characterize the users based on their actions. We denote the privacy concern of the users who are indifferent between using or not using, that is users with $E(U^u(v,r)) = E(U^n(v,r))$, as $\tilde{v} = X/sE(m)$. It can be seen that these indifferent users' privacy concern does not depend on their IT illiteracy r. We denote the privacy concern of users who are indifferent between using the website or expending effort to discover the website type, that is users with $E(U^u(v,r)) = E(U^e(v,r))$, as $\check{V}(r) = [[1-\theta]X + rE(\eta)]/[1-\theta]sm_H$. Finally, we denote the privacy concern of users who are indifferent between expending effort to discover the website type or not using the website, that is users with $E(U^e(v,r)) = E(U^n(v,r))$ as $\widehat{V}(r) = [\theta X - rE(\eta)]/\theta sm_L$. Note that the privacy concern of the users who are indifferent between using or expending effort ($\check{V}$) and not using or expending effort ($\widehat{V}$) characterized above, depends on their IT illiteracy (r). Therefore, $\check{V}$ and $\widehat{V}$ are functions of r and can be shown as indifference lines. Where the lines $v = \tilde{v}$, $\check{V}$, and $\widehat{V}$ intersect, we have $r = \tilde{r} = \theta[1-\theta]X(m_H - m_L)]/E(m)E(\eta)$. The indifference lines and their intersection points are illustrated in Figure 3.

Figure 3 Users' Expected Utility


r
$\widehat{V}$
$\check{V}$
$E(U^u) > E(U^n) > E(U^e)$
$E(U^n) > E(U^u) > E(U^e)$
VI
I
$E(U^u) > E(U^e) > E(U^n)$
$E(U^n) > E(U^e) > E(U^u)$
$\tilde{r}$
V
II
IV
III
$E(U^e) > E(U^u) > E(U^n)$
$E(U^e) > E(U^n) > E(U^u)$
$\underline{v}$
$\frac{X}{sm_H}$
$\tilde{v}$
$\frac{X}{sm_L}$
$\overline{v}$
v

Users' optimal decision can be represented as their best response to the website's level of obfuscation given the users' belief about the website type. Each user chooses the action $K_{v,r} \in \{u, e, n\}$ which maximizes her expected utility. For example, a given user uses the website if her expected utility of using the website is larger than the expected utility of both not using the website and expending effort to discover the website type. Accordingly, the users' best response function is defined as $B_{v,r}(\eta_L,\eta_H)=\underset{K_{v,r}}{\operatorname{argmax}}E(U^{K_{v,r}}(v,r))$. Comparing the expected utilities in (2), we can write the users' best response function as:

$$B_{v,r}(\eta_L,\eta_H)=\begin{cases} u, & \text{if} \quad v\leqslant\tilde{v} \ \& \ v\leqslant\check{V}(r) \\ e, & \text{if} \quad \check{V}(r)<v<\widehat{V}(r) \\ n, & \text{if} \quad v>\tilde{v} \ \& \ v\geqslant\widehat{V}(r) \end{cases} \tag{3}$$

As shown in Figure 3, in regions I and II the expected utility of not using the website, $E(U^n(v,r))$, is higher than both the expected utility of using the website, $E(U^u(v,r))$, and the expected utility of expending effort to discover the website type, $E(U^e(v,r))$. Therefore, the optimal decision for the users in these regions is to not use the website corresponding to $B_{v,r}(\eta_L,\eta_H)=n$. In regions III and IV the expected utility of expending effort to discover the website type is higher than both the expected utility of using and the expected utility of not using the website. Therefore, the optimal decision for the users in these regions is to expend effort to discover the website type, corresponding to $B_{v,r}(\eta_L,\eta_H)=e$. In regions V and VI the expected utility of using the website is higher than the expected utility of not using the website and the expected utility of expending effort to discover the website type. Therefore, the optimal decision for the users in these regions is to use the website, corresponding to $B_{v,r}(\eta_L,\eta_H)=u$.

3.1. Demand Specification

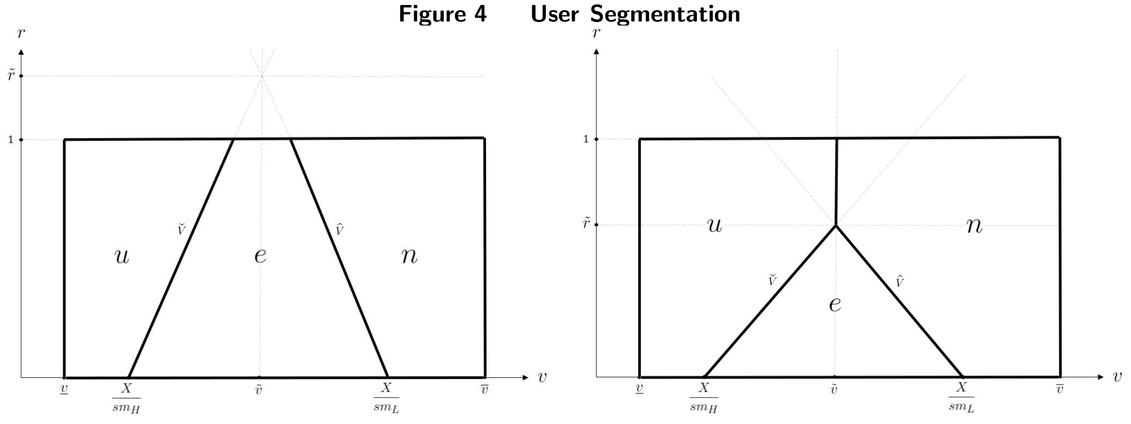
As obfuscation $(E(\eta))^3$ increases, the indifference lines between using and expending effort (the slope of $\check{V}$ with respect to r is $E(\eta)/[1-\theta]m_H$) and between expending effort and not

[3] Hereafter, we refer to expected obfuscation (composed of two possible website types' level of obfuscation) simply as obfuscation, as these two concepts are closely related. As we show in our analysis, a low-type website sets obfuscation to zero, and therefore, expected obfuscation in equilibrium is perfectly correlated with a high-type website's level of obfuscation.

using (the slope of $\widehat{V}$ with respect to r is $-E(\eta)/\theta m_L$) become more steep. Therefore, the crossing point of these two lines ($\tilde{r}$) moves down. We can specify demand depending on the obfuscation as compared to the threshold $\tilde{\eta} = \theta[1-\theta]X[m_H - m_L]/E(m)$. We discuss the significance of $\tilde{\eta}$ later in our analysis.

Figure 4 shows the user segments for the different expected levels of obfuscation. Where obfuscation is low ($E(\eta) < \tilde{\eta}$, Figure 4.(a)), the crossing point between $\check{V}$ and $\widehat{V}$ occurs above 1 ($\tilde{r} > 1$). In this case, there exist some users at any level of IT illiteracy r that have the incentive to expend effort to discover website type. As obfuscation increases to $E(\eta) = \tilde{\eta}$, the crossing point occurs at 1 ($\tilde{r} = 1$). In this case, the most IT illiterate user ($r = 1$) with an average privacy concern ($v = \tilde{v}$) is indifferent between using the website, not using the website, and expending effort to discover the website type. The threshold $\tilde{\eta}$ is the obfuscation level for which this special case occurs. As obfuscation increases even more ($E(\eta) > \tilde{\eta}$), the crossing point occurs at below 1 ($\tilde{r} < 1$). In this case, users with $\tilde{r} \leqslant r \leqslant 1$ do not have an incentive to expend effort to discover website type, and either use or not use the website irrespective of their privacy concern v.

Figure 4 User Segmentation



(a) $E(\eta) < \tilde{\eta}$ **(b)** $E(\eta) \geq \tilde{\eta}$

As explained above, only those users expend effort who know that given the cost of obfuscation, they gain positive utility from the website if they discover it to be low-type.

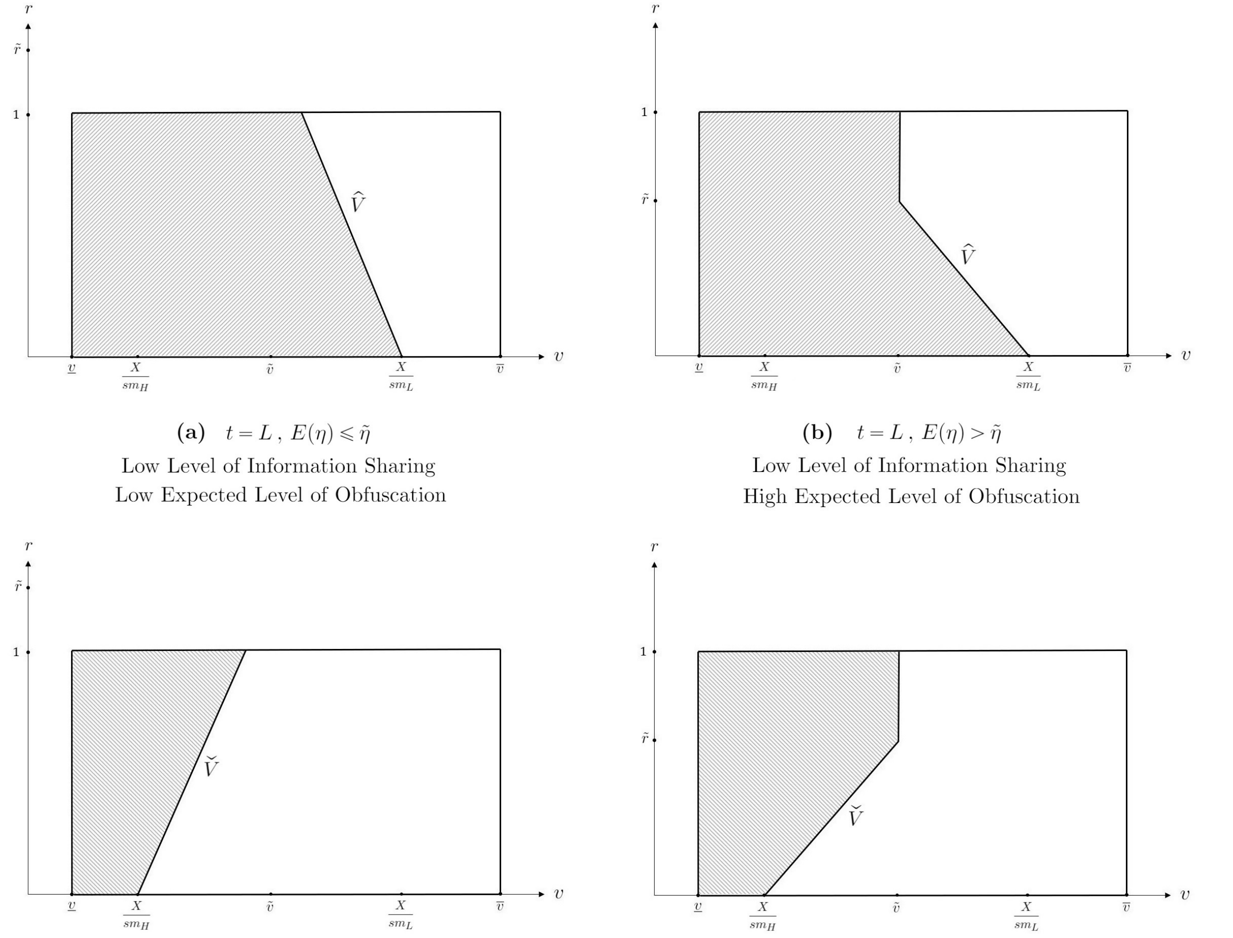
Therefore, users who expend effort, use the website only if the website is low-type, which implies that a low-type website's demand is composed of users who use the website without expending effort, and users who expend effort. The high-type website's demand, on the other hand, is composed only of users who use the website without expending effort, because if users expend effort and discover the true level of information sharing to be high, then they do not use the website. Given the users' best response function and the expected level of obfuscation, demand for each website type is characterized as:

$$\begin{aligned} D_L(\eta_L,\eta_H) &= \frac{\int_0^{\min\{\tilde{r},1\}} \widehat{V}\,dr + \int_{\min\{\tilde{r},1\}}^1 \tilde{v}\,dr - \underline{v}}{\overline{v}-\underline{v}} \\ D_H(\eta_L,\eta_H) &= \frac{\int_0^{\min\{\tilde{r},1\}} \check{V}\,dr + \int_{\min\{\tilde{r},1\}}^1 \tilde{v}\,dr - \underline{v}}{\overline{v}-\underline{v}} \end{aligned} \tag{4}$$

Considering the demand, the effect of obfuscation is that it makes it harder for users to discover the type of website, that is, it reduces users who expend effort. This shifts some of the users to not use the website, and others to use the website, both without expending effort. This increases the ratio of users who use the website, which improves the demand for the high-type website. On the other hand, this effect reduces the total ratio of users who either use the website or expend effort, which constitutes the low-type website's demand. The demand for each type of website is illustrated in Figure 5.

Because obfuscation impacts demand only through users who expend effort, users with $r > \tilde{r}$ are not impacted by obfuscation. This diminishes the impact of obfuscation on demand beyond $\tilde{\eta}$. Therefore, the impact of obfuscation on demand has two distinct pieces. Where obfuscation is low ($E(\eta) < \tilde{\eta}$, Figures 5.(a) and 5.(c)), obfuscation impacts the demand for users with any level of IT illiteracy ($r \leqslant 1$). However, where obfuscation is high ($E(\eta) > \tilde{\eta}$, Figures 5.(b) and 5.(d)), obfuscation impacts the demand only for IT literate users (users with $r < \tilde{r}$) who expend effort to discover website type. Therefore, in the second

Figure 5 Demand for Each Type of Website



(a) $t = L$, $E(\eta) \leqslant \tilde{\eta}$
Low Level of Information Sharing
Low Expected Level of Obfuscation

(b) $t = L$, $E(\eta) > \tilde{\eta}$
Low Level of Information Sharing
High Expected Level of Obfuscation

(c) $t = H$, $E(\eta) \leqslant \tilde{\eta}$
High Level of Information Sharing
Low Expected Level of Obfuscation

(d) $t = H$, $E(\eta) > \tilde{\eta}$
High Level of Information Sharing
High Expected Level of Obfuscation

piece where obfuscation is high, its impact on demand is diminished. This phenomenon has an important implication for the optimal level of obfuscation, as it makes the demand function piece-wise with respect to obfuscation (two pieces: $E(\eta) < \tilde{\eta}$ and $E(\eta) > \tilde{\eta}$). As we explain in the next section (Section 3.2), in the case where an interior solution exists, this drives the optimal level of obfuscation: in this case a high-type website increases the obfuscation beyond $\tilde{\eta}$, after which the marginal return of obfuscation is dampened, and the optimal level of obfuscation is reached.

3.2. Bayesian Equilibrium Analysis

As discussed, the website provides a free service to users, but charges third parties a fixed and exogenous price per user, which we normalize to 1. Therefore, website's revenue is given as $D_t(\eta_L, \eta_H)m_t$. The obfuscation cost function is $C(\eta_t)$, where $C'(\eta_t) > 0$, $C''(\eta_t) \geqslant 0$ and $C(0) = 0$. These standard assumptions imply that obfuscation is costly and that the website chooses easy obfuscation practices first, which makes obfuscation increasingly difficult. In other words, as website employs additional obfuscation practices, it becomes increasingly difficult to find new ways to prevent users from discovering the website type. We also allow for a linear cost function (where $C''(\eta_t) = 0$) to extend our results to this special case.

The website's profit function is given as:

$$\Pi_t = D_t(\eta_L, \eta_H)m_t - C(\eta_t) \tag{5}$$

Note that the website revenue is composed of the demand (D_t) times the level of information sharing (m_t, which equates to the per-user revenue). Content sensitivity and user privacy concern impact the website profit only indirectly through the demand. On the other hand, the level of information sharing (m_t) impacts revenue both directly as the per-unit revenue, and indirectly through demand (D_t), as users may refrain from using a website that extensively shares their data with third-parties.

The optimal decision of a type t website can be represented as its best response to users decisions. The website maximizes its profit function by choosing the level of obfuscation η_t. The best response function of a type t website in reaction to the collective set of all individual user actions $\{K_{v,r} | \underline{v} \leqslant v \leqslant \overline{v}\,,\, 0 \leqslant r \leqslant 1\}$ is therefore given as:

$$B_t(\{K_{v,r} | \underline{v} \leqslant v \leqslant \overline{v}\,,\, 0 \leqslant r \leqslant 1\}) = \underset{\eta_t}{\operatorname{argmax}}\, \Pi_t \tag{6}$$

The Bayesian Nash equilibria (Harsanyi 1968) of this game are vectors $(\eta_t^*, K_{v,r}^*)$ such that:

$$\begin{aligned} B_{v,r}(\eta_L^*, \eta_H^*) = K_{v,r}^* \quad , \qquad \forall \ (v \in [\underline{v}, \overline{v}] \ , \ r \in [0,1]) \\ B_t(\{K_{v,r}^* | \ \underline{v} \leqslant v \leqslant \overline{v} \ , \ 0 \leqslant r \leqslant 1\}) = \eta_t^* \quad , \qquad \forall \ t \in \{L, H\} \end{aligned} \tag{7}$$

Considering the impact of obfuscation on website profit in (5), it can be seen that for a low-type website, there is no incentive to increase obfuscation, as it both decreases demand and increases the costs. However, if the positive effect of obfuscation on demand outsizes the negative effect of increased costs, a high-type website has an incentive to obfuscate. Our first lemma characterizes the equilibrium obfuscation level. The details for deriving the equilibrium and all proofs are relegated to the Appendix.

LEMMA 1. **Level of Obfuscation in Equilibrium**) *A low-type website does not obfuscate ($\eta_L^* = 0$).) Where the marginal cost of obfuscation at the point $\tilde{\eta}/[1-\theta]$ is low ($C'(\tilde{\eta}/[1-\theta]) < 1/2s[\overline{v} - \underline{v}]$), there is an interior solution for the high-type website's optimal level of obfuscation, where $\eta_H^* > \tilde{\eta}/[1-\theta]$.) Where the marginal cost of obfuscation at the point $\tilde{\eta}/[1-\theta]$ is high ($C'(\tilde{\eta}/[1-\theta]) \geqslant 1/2s[\overline{v} - \underline{v}]$), the high-type website's optimal level of obfuscation depends on the curvature of the obfuscation cost function $C''(\eta_H)$. If the cost of obfuscation is convex ($C''(\eta_t) > 0$), then there is an interior solution for the high-type website's optimal level of obfuscation, where $0 < \eta_H^* \leqslant \tilde{\eta}/[1-\theta]$. If the cost of obfuscation is linear ($C''(\eta_t) = 0$), then there is a corner solution for the high-type website's optimal level of obfuscation, where it does not obfuscate ($\eta_H^* = 0$).*

We find that a low-type website does not obfuscate. This is intuitive, because obfuscation is costly, and it only reduces the website's demand if it is a low-type. Where the cost of obfuscation is convex (obfuscation becomes increasing difficult and costly as the website obfuscates more), the high-type website obfuscates enough that for users with high level of

IT illiteracy, it is not optimal to expend effort to discover the website type. As previously explained, this occurs only for high levels of obfuscation ($E(\eta) > \tilde{\eta}$, Figure 4.(b)). Where the cost of obfuscation is linear or slightly concave, if the slope of the function is high, then the high-type website does not obfuscate. If the slope of the function is low, however, then a high-type website obfuscates similar to the previous case. Where the cost of obfuscation is strongly concave, if the slope of the function is high, then a high-type website does not obfuscate, as it is too costly to do so. If the slope of the function is low, then a high-type website obfuscates and the obfuscation level is in a way that even users with very high level of IT illiteracy ($r > \tilde{r}$) expend effort to discover the website type.

Next, we use comparative statics to study the impact of content sensitivity on obfuscation.

Proposition 1. **Content Sensitivity and Obfuscation**) *A low-type website does not obfuscate, irrespective of its content sensitivity.) If the cost of obfuscation is linear and the content sensitivity is high* ($s \geqslant \frac{1}{2C'(\frac{\tilde{\eta}}{1-\theta})[\overline{v}-\underline{v}]}$)*, then a high-type website does not obfuscate* ($\eta^*_H = 0$)*. Otherwise, a high-type website obfuscates* ($\eta^*_H > 0$) *and this level of obfuscation is decreasing in its content sensitivity* ($\partial \eta^*_H / \partial s < 0$)*.*

To consider the impact of content sensitivity on obfuscation, note that the low-type website's demand is composed of users who use the website without expending effort, and users who expend effort. The high-type website's demand, on the other hand, is composed only of users who use the website without expending effort, because if users expend effort and discover the true level of information sharing to be high, then they do not use the website. For high-type websites, obfuscation increases demand by increasing the ratio of users who do not expend effort. High content sensitivity, however, makes it more valuable for users to discover the website type, thereby mitigating the effect of obfuscation on demand ($\partial^2 D_t/\partial\eta_t\partial s < 0$). Therefore, as content sensitivity increases, obfuscation yields diminishing benefits, resulting in website choosing a lower level of obfuscation. Particularly, where cost of obfuscation is linear or concave, if content sensitivity is sufficiently high, the website does not obfuscate.

PROPOSITION 2. **Website Value and Obfuscation**) *A low-type website does not obfuscate, irrespective of the website value.) The high-type website's level of obfuscation is increasing in the website value for all* $X < \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})E(m)}{\theta[m_H-m_L]}$. *Increasing* X *beyond* $\frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})E(m)}{\theta[m_H-m_L]}$ *does not impact the level of obfuscation.*

As discussed above, obfuscation increases the demand of high-type websites by increasing the proportion of users who do not expend effort. While $X < \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})E(m)}{\theta[m_H-m_L]}$, that is $\eta_H^* > \frac{\tilde{\eta}}{1-\theta}$ (Figure 5(d)), as the website value increases the proportion of users whom their decisions are affected by obfuscation increases ($\tilde{r}$ increases). This intensifies the effect of obfuscation on demand, whereby $\partial^2 D_t/\partial\eta_t\partial X > 0$. Therefore, when $\eta_H^* > \frac{\tilde{\eta}}{1-\theta}$ as website value X increases, obfuscation yields higher benefits for a high-type website, resulting in these website choosing a higher level of obfuscation. When $X \geqslant \frac{C'^{-1}(\frac{1}{2s[\overline{v}-\underline{v}]})E(m)}{\theta[m_H-m_L]}$, that is $\eta_H^* \leqslant \frac{\tilde{\eta}}{1-\theta}$ (Figure 5(c)), the proportion of users whom their decisions are affected by obfuscation is independent of the website value. Therefore, when $\eta_H^* \leqslant \frac{\tilde{\eta}}{1-\theta}$ increasing the website value does not affect the impact of obfuscation on the website's profit.

4. Empirical Analysis

In this section, we present our empirical analysis of the website obfuscation of third-party sharing. This analysis provides some empirical support for our analytical results. We examine websites' reaction to a form of explicit user privacy request, known as Do Not Track (DNT), to analyze the level of obfuscation. Users utilize passive privacy tools, which signal to the website that such users are either concerned about their privacy, concerned about being tracked, or both. Khatibloo et al. (2018) reveal that 25% of American adults use the Do Not Track browser option to protect their privacy. Although there is generally no legal requirement in Canada (where our study is conducted) to honor a user's DNT request, it does send a clear message about the user request for privacy.

A website has three possible reactions to a DNT request. Two of these options are straightforward: it can either respect the request by reducing the number of third-parties or abuse the request by increasing the number of third-parties employed. The third option is to both add and drop third-parties. This option creates confusion in that it confuses privacy and third-party monitoring services about whether the website is complying with the DNT request or not. This is especially true for privacy-violating third-parties, i.e. those third-parties that are known to track users and collect their information. We use this latter website reaction as a measure obfuscation, as it increases the effort needed for users to determine the behavior of the website.

4.1. Data and Measurements

Alexa Internet[4] provides rankings for publisher websites within different website subject categories. From these categories, we take the list of the 500 most-visited websites from two categories: News and Health. These two categories were selected with the intention of finding website subject categories for which users have different privacy concerns and intentions to disclose personal information and browsing behavior due to the nature of the subject content. Health information is generally thought to have higher content sensitivity than News. Many websites allow users to create accounts and profiles. This is particularly important for Health websites, for example `webmd.com`, `mayoclinic.org`, `medscape.com`, `drugs.com`, `psychologytoday.com`. The third-parties used by these website have access

[4] Alexa.com

to the detailed health-related information of users. The Electronic Frontier Foundation provides some evidence for this, where a Health website shared users' health related data with third-parties (Quintin 2015). This makes the third-party sharing behavior of Health website much more sensitive than News websites. Even though not all websites collect user-specific data, user behavior on Health websites can still be sensitive, as it can be used to predict health-related facts about a user. For example, consider a user looking at a heart condition page on `cdc.gov` or `heart.org`. Such a user can be predicted to be of high risk for heart conditions by third-party trackers. Use of third-party cookies enables trackers to collect such information and depict a profile of the user with the information on their behavior in other websites. Therefore, compared to News websites, Health websites are more sensitive in terms of privacy concern. This line of reasoning is consistent with previous studies on website categories and content sensitivity, including Gopal et al. (2018).

After removing duplicate websites, the list of the 500 most-visited websites is reduced to 480 News and 441 Health websites. We also removed websites for which the average number of monthly users was unknown, resulting in a total of 676 websites (391 News and 285 Health). These were removed to limit our analysis to websites for which full data is available. We then narrowed our set of websites to a randomly selected balanced list of 200 from each category for our study to make data collection more manageable. We used Lightbeam (an add-on for Firefox) to record the connections between third-parties and websites. An automated browser accessed the homepage of each website for a particular category (that is, News or Health) with 11 runs through the entire list of websites within each category. For each of the 11 runs, the list of websites was randomly sequenced. All cookies and cached data were cleared before the first of 11 runs for each category. We removed the first run from our analysis, as this is primarily when cookies are initialized (Gopal et al. 2018). For each website, we collected HTML source code and all connections made between the website and third-parties within the first 5 seconds of visiting the website. The same process was repeated with the DNT request turned on within the browser settings.

We distinguish two types of connections between websites and third-parties. The first type is connections that are requested in the page's HTML source code, which we label as primary connections. The second type are the ones that have not been requested in the page's HTML source code, but which are redirected to other third-parties. We term such connections as secondary. We observed instances in which a particular third-party served as a primary third-party in one user visit, and a secondary third-party in another visit for the same website. We also separated third-parties based on whether they collect user information and track users' profiles. This is based on the list of tracking third-parties from EasyPrivacy[5], which is a list of third-parties that track user activity online.

We measure content sensitivity by websites category (Health has high content sensitivity and News has low content sensitivity) and use Alexa Internet ranking as a proxy for website value. Our main independent variables are $Category \in \{0,1\}$ which denotes the website category, 0 for News and 1 for Health, and $Rank$ which denotes the Alexa Internet ranking for websites. Other than website content sensitivity and value, several factors in the website environment such as website's business model and required functionalities may affect their information sharing behavior. We use two variables to control for the effect of these factors. First, we consider whether the website uses advertising as a source of income to be a control variable. This is a binary variable denoted as 1 if the website uses advertising, labelled as Advertising. Second, we consider whether the website utilizes a

[5] https://easylist.to/

user login option as a control variable. This is another binary variable denoted as 1 if the website has a login option, labelled as Login. Both variables were manually determined for each website. Because the variables Advertising (whether there is an advertising banner on the page) and Login (whether there is a login button on the page) are objectively observable, only one rater was utilized to determine the data values.

We use a multi-stage model that accounts for other managerial and economic factors. We only consider privacy-violating third-parties, as these are expected to have the negative privacy costs as perceived by users. Additionally, because websites do not have full control on the number of secondary third-parties, we only consider primary third-parties in our analysis. The websites in our study used 17,965 primary third-parties in total, where 58% of these third-parties are privacy-violating. A summary of the empirical model notation is shown in Table 3. Descriptive statistics and the and correlation matrix are provided in Tables 4 and 5, respectively.

4.2. Estimation Approach and Econometric Considerations

We examine the reaction of websites to an explicit user privacy request through DNT, to capture the websites' level of obfuscation. In response to DNT, websites have three options: respect, abuse without obfuscation, and abuse with obfuscation. Figure 6 shows the websites decision process, regarding their reaction to DNT. The decision process consists of two stages. The first stage is the choice between abusing and not abusing (respecting) the user privacy request. The second stage is a choice between obfuscate and not obfuscate for those websites who have chosen to abuse user privacy request in the first stage. To analyze this process, we use the sequential logit model proposed by Mare (1980).

The sequential logit model predicts the probability of website's decisions in each stage by estimating a logistic regression. In the first stage, the sequential logit model estimates

Table 3 Empirical Model Notation

Notation	Definition
$Category$	A proxy for website content sensitivity, where $Category \in \{0,1\}$, 0 for News and 1 for Health (this captures s in the analytical model).
$Rank$	A proxy for website value (this captures X in the analytical model).
$Advertising$	Whether the website displays advertisements, where $Advertising \in \{0,1\}$ denoted as 1 if the website uses advertising, and 0 otherwise.
$Login$	Whether the website utilizes a user login option, where $Login \in \{0,1\}$ denoted as 1 if the website has a login option, and 0 otherwise.
$Primary$	Number of privacy-violating primary third-parties connected to the website.
$Size$	Average number of monthly users.
$Abuse$	Website decision in response to user privacy request, where $Abuse \in \{0,1\}$ denoted as 1 if the website decides to increase the number of third-parties in response to the user privacy request, and 0 otherwise.
$Abuse - Level$	Number of third-parties added in response to user privacy request.
$Obfuscate$	Website decision with respect to the obfuscation of information sharing with third-party, where $Obfuscate \in \{0,1\}$ denoted as 1 if the website decides to obfuscate, and 0 otherwise (this captures whether η_t equals zero or is greater than zero in the analytical model).
$Obfuscation - Level$	Website level of obfuscation (this captures η_t in the analytical model).

Table 4 Descriptive Statistics

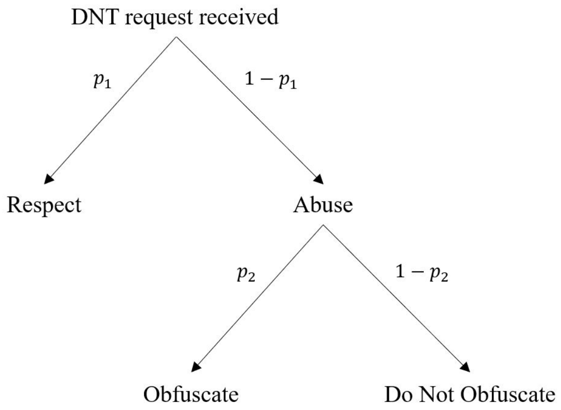
Statistic	N	Mean	St. Dev.	Min	First Quartile	Third Quartile	Max
Category	400	0.500	0.501	0	0	1	1
Rank	400	193.062	119.130	2	90.8	284.2	434
Advertising	400	0.557	0.497	0	0	1	1
Login	400	0.728	0.446	0	0	1	1
Primary	400	9.392	5.335	0	5.8	13	29
Abuse	400	0.232	0.423	0	0	0	1
Abuse-Level	400	0.355	0.872	0	0	0	9
Obfuscate	400	0.055	0.228	0	0	0	1
Obfuscation-Level	400	0.458	1.047	0	0	0	8

$p_1 = P(Abuse|X)$, using the overall dataset, where X is the vector of explanatory variables. In the second stage, the sequential logit model estimates $p_2 = P(Obfuscate|X, Abuse)$, using the sub-sample of websites that did not chose to respect user privacy request in the first stage. Let J be the set of all websites, and $j \in J$ represents each website. The logit

Table 5 Correlation Matrix

	Category	Rank	Advertising	Login	Primary	Abuse	Abuse-Level	Obfuscate	Obfuscation-Level
Category	1	-0.102	-0.680	-0.275	-0.521	-0.302	-0.258	-0.197	-0.294
Rank	-0.102	1	0.060	0.107	-0.005	0.006	-0.027	-0.120	-0.077
Advertising	-0.680	0.060	1	0.291	0.524	0.324	0.271	0.215	0.313
Login	-0.275	0.107	0.291	1	0.334	0.137	0.114	0.049	0.112
Primary	-0.521	-0.005	0.524	0.334	1	0.247	0.145	0.211	0.213
Abuse	-0.302	0.006	0.324	0.137	0.247	1	0.740	0.438	0.795
Abuse-Level	-0.258	-0.027	0.271	0.114	0.145	0.740	1	0.191	0.870
Obfuscate	-0.197	-0.120	0.215	0.049	0.211	0.438	0.191	1	0.608
Obfuscation-Level	-0.294	-0.077	0.313	0.112	0.213	0.795	0.870	0.608	1

Figure 6 Website Decision Process



regression models for the first and second stages are as follows:

$$Abuse_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 Advertising_j + \beta_4 Login_j + \epsilon_{1j} \tag{8}$$

$$Obfuscate_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 Advertising_j + \beta_4 Login_j + \epsilon_{2j} \tag{9}$$

In the first stage model, the dependent variable *Abuse* is a binary variable which takes the value 1 if the website increases the number of third-parties in response to the user privacy request in the first stage, and 0 otherwise. In the second stage model the dependent variable $Obfuscate$ is a binary variable which takes the value 1 if the website both adds and drops third-parties in the second stage, and 0 otherwise. There is a high correlation between websites' *Category* and *Advertising* $(-0.679, p < 0.001)$, thus if we use *Advertising* directly as a control variable, we expect to encounter multicollinearity issues.

To check the overall model for multicollinearity issues, we conduct the Farrar-Glauber test and the Theil test. The Farrar Chi-square values are 292.89 for the first model, and 46.70 for the second model. The Farrar Chi-squares are highly significant, implying that multicollinearity is present in both model specifications. The Theil test results for multicollinearity are 0.7084 (confirming multicollinearity) for the first model and 0.3980 (not confirming multicollinearity) for the second model. Next, we seek to locate specifically where the multicollinearity is located, and the results are presented in Table 6. We use the variance inflation factor (VIF), Farrar-Glauber F-test (Wi), and Klein test to locate the source of multicollinearity. the Farrar-Glauber F-test and Klein test both provide strong results for *Category* and *Advertising* being the source of the multicollinearity. While the VIF test does not provide proof of multicollinearity by itself for either model, the results are stronger for *Category* and *Advertising*. Collectively, we take this as evidence for the need to correct for multicollinearity between *Category* and *Advertising*.

Table 6 Test for Multicollinearity

	First Stage			Second Stage		
	VIF	Wi	Klein	*VIF*	Wi	Klein
Category	1.890	117.578	1	1.597	17.717	1
Rank	1.018	2.373	0	1.062	1.864	0
Advertising	1.900	118.872	1	1.482	14.326	1
Login	1.114	15.077	0	1.102	3.025	0

To deal with this problem, we develop a multi-stage model. First, we predict *Advertising* using the *Category* and all other control variables. Then we use the residuals from the model to predict the dependent variables for subsequent models. The first stage logistic regression model is:

$$Advertising_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 Login_j + \epsilon_{3j} \tag{10}$$

The coefficient for $Category$ is -3.408 and the effect of $Category$ on $Advertising$ is significant ($p < 0.01$). We take the residuals of this model and use it as a variable in the main model, denoted as $Ad_{res} = \epsilon_3$. This new variable captures all other variables that affect $Advertising$, except $Category$, $Rank$ and $Login$. In the next regression models, we use Ad_{res} as a control variable.

Unobservable variables include managerial, economic, and other factors, may impact the website decision regarding third-party usage. To resolve the omitted variable bias problem, we use the residuals from the following model ($Prim_{res} = \epsilon_{4j}$), which captures other factors that affect websites decisions, as another control variable in our main models. Our dependent variable ($Primary_j$) is the number of primary third-parties. Because this is a count value, we use negative binomial regression model which is appropriate for modeling count values. We also control for the total number of third-parties using an offset.

$$Primary_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 \epsilon_{3j} + \beta_4 Login_j + \epsilon_{4j} \tag{11}$$

Finally, to avoid multicollinearity issues and omitted variable bias, we estimate the following two logit regression models.

$$Abuse_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 \epsilon_{3j} + \beta_4 Login_j + \beta_5 \epsilon_{4j} + \epsilon_{5j} \tag{12}$$

$$Obfuscate_j = \beta_0 + \beta_1 Category_j + \beta_2 Rank_j + \beta_3 \epsilon_{3j} + \beta_4 Login_j + \beta_5 \epsilon_{4j} + \epsilon_{6j} \tag{13}$$

We also examine the magnitude of websites' reaction to user privacy request. We capture this using two additional dependent variables, $Abuse - Level$ and $Obfuscation - Level$. $Abuse - Level$ is the number of third-parties added in response to user privacy request, and $Obfuscation - Level$ is the summation of both the number of third-parties added and dropped with the user privacy request. Because these additional dependent variables are count values, we analyze them using negative binomial regression model.

The results for the first and second stage models are provided in Table 7. The effect of $Category$ on both $Abuse$ and $Abuse-Level$ in the first stage is significant and the coefficient is negative. This result shows that websites with less sensitive content are more likely to abuse user privacy request and the extent of abuse is also higher for the websites with less sensitive content. In the second stage the impact of $Category$ on $Obfuscate$ is significant and the coefficient is negative, implying that websites with low content sensitivity are more likely to obfuscate. This result is consistent with the findings in Proposition 1, where the obfuscation cost function is linear. The impact of $Category$ on $Obfuscation-Level$ is also significant in the negative binomial model and the coefficient is negative, which indicates that among the websites that obfuscate, the level of obfuscation of websites with low content sensitivity is higher. This result is consistent with the findings for both linear and convex obfuscation cost functions in Proposition 1. The impact of $Rank$ on $Obfuscate$ is significant and the coefficient is negative (-0.007), which indicates that less prominent websites are less likely to obfuscate. The impact of $Rank$ on $Obfuscation-Level$ is also negative and significant in the negative binomial model. This implies that among the websites that obfuscate, more prominent websites employ higher levels of obfuscation. These findings are in line with Proposition 2.

4.3. Robustness Check

To further investigate the robustness of our results, we repeat our analysis using both Primary and Secondary (Total) third-parties. The results of this analysis (as shown in Table 8) are fairly consistent with our previous findings. ~~It can be seen that considering all third-parties (both primary and secondary), user data from Health websites is additionally shared with more third-parties than News websites in response to a DNT request. Because we already know that Health websites reduce their number of primary third-parties more~~

Table 7 Website Reaction to DNT (Two-stage Model) Using Primary Third-Parties

	First Stage		Second Stage	
	Abuse	*Abuse − Level*	*Obfuscate*	*Obfuscation − Level*
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
Category	−1.542*** (0.290)	−1.530*** (0.272)	−2.656* (1.241)	−0.585*** (0.217)
Rank	0.000 (0.001)	−0.001 (0.001)	−0.007*** (0.002)	−0.001*** (0.000)
Ad_{res}	0.453*** (0.144)	0.421*** (0.135)	−0.987 (0.673)	0.141 (0.096)
Login	0.398 (0.336)	0.500 (0.316)	0.048 (0.807)	0.091 (0.217)
$Prim_{res}$	0.327 (0.344)	−0.100 (0.319)	1.336 (1.177)	−0.241 (0.250)
Constant	−0.736* (0.394)	−0.756** (0.365)	0.195 (0.979)	0.946*** (0.231)
Observations	400	400	93	93
AIC	393.349	570.169	94.972	294.481
Hosmer-Lemeshow p-value	0.599	0	0.439	1.000
McFadden's R^2	0.121	0.082	0.184	0.051
Cox & Snell R^2	0.123	0.117	0.183	0.151
Cameron & Windmeijer R^2	0.121	0.191	0.184	0.225

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

~~than News websites in response to DNT (Table 7), this implies that the number of secondary third-parties increases more for Health websites compared to News websites in response to DNT. This may be due to the Health websites being less cognizant of how third-parties share data with secondary third-parties, which enables both primary and secondary third-parties to additionally share user data from health websites with more secondary third-parties compared to News websites in response to DNT. Irrespective of the change in the level of information sharing (number of third-parties) the impact of content sensitivity on the website strategic obfuscation remains in line with our model. Consistent with the model using only primary third-parties, we find that the impact of increased content-sensitivity is decreased obfuscation, which adds to the robustness of this finding.~~ The impact of *Rank* on *Obfuscate* is no longer significant, as the decision to share information with secondary third-parties is not directly made by the website.

Table 8 **Website Reaction to DNT (Two-stage Model) Using Total (Primary+Secondary) Third-Parties**

	First Stage		Second Stage	
	$Abuse$	$Abuse-Level$	$Obfuscate$	$Obfuscation-Level$
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
$Category$	-0.674***	-1.512***	-2.591***	-1.305***
	(0.246)	(0.275)	(0.599)	(0.145)
$Rank$	0.001	-0.001	-0.001	-0.001**
	(0.001)	(0.001)	(0.002)	(0.0005)
Ad_{res}	0.504***	0.563***	1.111***	0.452***
	(0.128)	(0.133)	(0.299)	(0.074)
$Login$	0.282	0.451	0.116	0.116
	(0.287)	(0.321)	(0.614)	(0.165)
$Total_{res}$	-0.024	0.586*	3.597***	1.149***
	(0.288)	(0.309)	(0.857)	(0.158)
Constant	-1.079***	0.761*	2.692***	3.007***
	(0.354)	(0.392)	(0.841)	(0.202)
Observations	400	400	93	93
AIC	462.831	1045.172	102.489	747.015
Hosmer-Lemeshow p-value	0.069	1.000	0.016	1.000
McFadden's R^2	0.064	0.042	0.377	0.113
Cox & Snell R^2	0.074	0.108	0.376	0.552
Cameron & Windmeijer R^2	0.064	0.171	0.377	0.537

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

We also employed the average number of monthly users ($Size$) to reflect $Rank$ as a robustness check. We replaced $Rank$ with $log(Size)$ in all models. Results are provided in Table 9 (for primary third-parties) and Table 10 (for total third-parties). It can be seen that both tables are consistent with the main results.

Moreover, we analyze the empirical results for estimating coefficients in equations (8) and (9) directly. Results of the one-stage models are provided in Table 11. The effect of $Category$ on $Abuse$ and $Obfuscate$ are significant, consistent with the main results (two-stage models). However the effect of $Category$ on abuse and obfuscation levels are not significant. The effect of $Rank$ on $Obfuscation-Level$ is significant, consistent with the main results. However the effect of $Rank$ on $Obfuscate$ is not significant. While some of the one-stage model results are different from two-stage models, the results for the main coefficients are consistent with the analytical model findings in Proposition 1 and Proposition 2.

Table 9 Website Reaction to DNT (Two-stage Model With log(Size)) Using Primary Third-Parties

	First Stage		Second Stage	
	Abuse	*Abuse − Level*	*Obfuscate*	*Obfuscation − Level*
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
Category	-1.442***	-1.387***	-2.178*	-0.453**
	(0.291)	(0.272)	(1.233)	(0.215)
log(*Size*)	0.190**	0.209***	0.589***	0.144***
	(0.088)	(0.075)	(0.185)	(0.046)
Ad_{res}	0.442***	0.407***	0.949	0.129
	(0.144)	(0.134)	(0.677)	(0.097)
Login	0.381	0.459	-0.166	0.039
	(0.339)	(0.314)	(0.834)	(0.216)
$Prim_{res}$	0.289	-0.023	1.518	-0.159
	(0.349)	(0.318)	(0.977)	(0.252)
Constant	-3.457***	-3.836***	-9.064***	-1.309*
	(1.237)	(1.073)	(2.747)	(0.681)
Observations	400	400	93	93
AIC	389.616	564.693	94.535	292.251
Hosmer-Lemeshow p-value	0.494	1.000	0.652	1.000
McFadden's R^2	0.129	0.092	0.189	0.059
Cox & Snell R^2	0.131	0.129	0.187	0.171
Cameron & Windmeijer R^2	0.129	0.210	0.188	0.259

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table 10 Website Reaction to DNT (Two-stage Model With log(Size)) Using Total Third-Parties

	First Stage		Second Stage	
	Abuse	*Abuse − Level*	*Obfuscate*	*Obfuscation − Level*
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)
Category	-0.715***	-1.452***	-2.506***	-1.252***
	(0.248)	(0.277)	(0.594)	(0.143)
log(*Size*)	-0.040	0.128	0.193	0.118***
	(0.084)	(0.094)	(0.204)	(0.044)
Ad_{res}	0.505***	0.560***	1.110***	0.448***
	(0.127)	(0.132)	(0.296)	(0.074)
Login	0.308	0.351	-0.034	0.038
	(0.286)	(0.318)	(0.652)	(0.168)
$Total_{res}$	-0.017	0.577*	3.490***	1.147***
	(0.286)	(0.309)	(0.847)	(0.155)
Constant	-0.327	-1.035	0.054	1.276**
	(1.161)	(1.307)	(2.612)	(0.580)
Observations	400	400	93	93
AIC	463.924	102.555	1043.890	745.036
Hosmer-Lemeshow p-value	0.675	0.020	1.000	1.000
McFadden's R^2	0.062	0.376	0.044	0.115
Cox & Snell R^2	0.072	0.376	0.111	0.560
Cameron & Windmeijer R^2	0.062	0.376	0.176	0.544

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

5. Concluding Remarks

With the increasing importance of user privacy, understanding the interaction of users, websites, and third-parties is more important than ever. In this paper, we focus on the

Table 11 Website Reaction to DNT (One-stage Model) Using Primary Third-Parties

	First Stage		Second Stage			
	Abuse	*Abuse − Level*	*Obfuscate*	*Obfuscation − Level*		
	(Logit)	(Negative Binomial)	(Logit)	(Negative Binomial)		
Category			-0.754**	-1.121	-0.744**	-0.279
			(0.346)	(0.891)	(0.322)	(0.232)
Rank			-0.001	-0.007***	-0.001	-0.002***
			(0.001)	(0.002)	(0.001)	(0.001)
Advertising			1.300***	17.353	1.305***	0.519*
			(0.386)	(1587.729)	(0.369)	(0.294)
Login			0.269	-0.116	0.325	0.033
			(0.337)	(0.741)	(0.315)	(0.211)
Constant			-1.866***	-16.860	-1.782***	0.552
			(0.504)	(1587.730)	(0.473)	(0.368)
Observations			400	400	93	93
AIC			391.651	566.100	92.946	292.385
Hosmer-Lemeshow p-value			0.537	0	0.718	1.000
McFadden's R^2			0.120	0.086	0.184	0.052
Cox & Snell R^2			0.122	0.122	0.183	0.152
Cameron & Windmeijer R^2			0.120	0.198	0.185	0.227

Note: Standard errors in parentheses.
*p<0.1; **p<0.05; ***p<0.01

implicit contract between users and third-parties with respect to how websites share user information, as well as their effort to obfuscate the use of third parties. Developing analytical models for theoretical development and empirical models for external validation, this study examines the websites' obfuscation of user information sharing with third-parties. Our analytical model examines the website's decision regarding the extent of obfuscation of information sharing, given user privacy concerns. Interestingly and counter-intuitively, websites reduce the level of obfuscation as user privacy concern increases. Moreover, prominent websites are more prone to obfuscate the sharing of user information with third-parties.

Our results have important managerial insights for websites concerned about user privacy and informed user consent. Obfuscation of information sharing impairs the user's ability to make informed consent with respect to the privacy risks of using a website. An important contribution of this work is to explain the websites' incentive to obfuscate user information sharing. This has several implications for website practices toward information sharing. First, if user privacy is substantially impacted by the sensitive nature of the website content, websites should refrain from excessive monetization and the temptation to obfuscate third-party sharing. Secondly, it is primarily the most popular websites have the latitude to obfuscate. Managers of relatively less popular websites with more sensitive content have a greater incentive to refrain from obfuscation. Going beyond our results, we see current examples of heads of large platforms being called before the U.S. Congress to explain abusive practices. Eventually, threats of anti-trust enforcement and reputational damage to the brand could exact a substantial penalty for what are likely to be the comparatively moderate benefits of obfuscation.

Additionally, our results have important implications for policy makers concerned with website exploitation of user information. Policy makers need to be aware of information asymmetry problems and user privacy implications of information sharing as they design policies. One area of active policy making is in the domain of data protection regulation, for example European Union's General Data Protection Regulation (GDPR) or California's Consumer Privacy Act (CCPA). GDPR and CCPA require websites to acquire user consent (allow the user to either opt-in or opt-out) before their information is shared with third-parties. Some may argue that such regulation might make the issue of obfuscation moot. However, this is not necessarily the case. For the same reasons that website privacy policies do not convey all the information to users, acquiring consent does not necessarily resolve the information asymmetry between websites and users. First, the list of third-parties that websites are mandated to provide to users are typically long, confusing, and ineffective. Having such a list does not resolve information asymmetry as users have to engage in additional effort regarding the source and credibility of third-parties. Similar to privacy policies, these provided lists are often designed and presented in ways to confuse users and can persuade them to make risky privacy choices.

To further investigate the impact of data protection policies on obfuscation, we examined third-parties from a small sample of websites under the GDPR legal environment when accessing using a European IP address. We collected information from the reported lists of third-party vendors provided by websites (required by GDPR), examined the HTML source code for links to third-parties, and collected information on third-parties with whom the websites communicated using Lightbeam. We find that websites continue to use secondary third-parties under GDPR. It is difficult to determine if websites properly report these third-parties in the provided list because the list only consists of vendors' company names, not the URLs. This makes it hard for users to accurately determine the third-parties that have access to their data. We also examined the sharing behavior with and without DNT and find that some websites add and/or drop third-parties from their reported list of third-parties in reaction to the DNT request. These observations imply that obfuscation continues to be present even under the GDPR legal regime.

Thus, we are skeptical that current GDPR legislation, in and of itself, can effectively solve the problem of obfuscation. The cat and mouse game between regulators and obfuscators is likely to continue with ever evolving technology. Perhaps standardized and required transparency is the regulator's only real tool. Standardized disclosures are required for a variety of consumer financial contracts and corporate financial statement disclosures, and the same concepts could be applied to the website level of information sharing. Fines could reasonably be imposed for disclosure violations.

Websites seem to be reacting to the market forces surrounding user privacy concern and their ability or inability to exploit users' lack of information through third-party obfuscation. We confirm that it is primarily the most prominent websites that obfuscate their use of third-parties. Our analysis illustrates that obfuscation is not simply a matter of ability related to size and resources, but rather is a form of strategic exploitation. It is the large and prominent websites that obfuscate in order to additionally monetize user information. Small websites could use these insights to inform their interaction with users in terms of privacy and obfuscation. Even though we assume that small websites would exploit the if they could, that should not prevent small websites from highlighting how honest they are compared to their more dominant competitors.

We provide a simple and elegant model to capture the nuances involved in the obfuscation of third-parties. That said, in reality, the level of information sharing is more likely determined at the same time as the level of obfuscation in a dynamic manner. This makes a dynamic model of incomplete information a useful avenue for future research on this topic. Future work would also benefit from more refined measures of content sensitivity. If we have a better measure for content sensitivity, then researchers would be able to utilize this robust measure of content sensitivity to expand the number of categories and further validate our findings. Much future work is possible regarding the measurement of content sensitivity.

Another area for future work includes connecting this research to the impact of search engine optimization on obfuscation and user privacy. It would be useful to examine the effect of competition on websites' strategic decision-making with respect to obfuscation and privacy violation. Additionally, future analysis could examine the use of secondary third-parties, such as third-party advertisers not directly called by the website, using a sequential principal-agent model. Even though our study provides important implications for policy-making, we do not analytically address the policy-makers' incentive to limit or prohibit the level of obfuscation that websites utilize. Such analysis can have provide additional insights to help with policy-making in future studies.

References

Acar G, Englehardt S, Narayanan A (2020) No boundaries: data exfiltration by third parties embedded on web pages. *Proceedings on Privacy Enhancing Technologies* 2020(4):220–238.

Akerlof G (1970) The market for lemons: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics* 84(3):488–500.

Akerlof GA, Shiller RJ (2015) *Phishing for phools: The economics of manipulation and deception* (Princeton University Press).

Akerlof GA, Shiller RJ (2016) Manipulation and deception as part of a phishing equilibrium. *Business Economics* 51(4):207–212.

Arrow KJ (1985) The economics of agency. Technical report, In J. W. Pratt R. J. Zeckhauser (Eds.), Principals and agents: The structure of business (pp. 37-51). Boston: Harvard Business School Press.

Bai X, Gopal R, Nunez M, Zhdanov D (2012) On the prevention of fraud and privacy exposure in process information flow. *INFORMS Journal on Computing* 24(3):416–432.

Bushee BJ, Gow ID, Taylor DJ (2018) Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research* 56(1):85–121.

Chellappa RK, Shivendu S (2007) An economic model of privacy: A property rights approach to regulatory choices for online personalization. *Journal of Management Information Systems* 24(3):193–225.

Cisco (2021) Cisco 2021 consumer privacy survey. `https://www.cisco.com/c/dam/en_us/about/doing_business/trust-center/docs/cisco-cybersecurity-series-2021-cps.pdf`, accessed: 2021-11-20.

Eisenhardt KM (1989) Agency theory: An assessment and review. *Academy of management review* 14(1):57–74.

Ellison G, Wolitzky A (2012) A search cost model of obfuscation. *The RAND Journal of Economics* 43(3):417–441.

Englehardt S, Narayanan A (2016) Online tracking: A 1-million-site measurement and analysis. *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 1388–1401.

Englehardt S, Reisman D, Eubank C, Zimmerman P, Mayer J, Narayanan A, Felten EW (2015) Cookies that give you away: The surveillance implications of web tracking. *Proceedings of the 24th International Conference on World Wide Web*, 289–299.

Goldstein A, Eaton C (2021) Asymmetry by design? identity obfuscation, reputational pressure, and consumer predation in us for-profit higher education. *American Sociological Review* 86(5):896–933.

Gopal RD, Hidaji H, Patterson RA, Rolland E, Zhdanov D (2018) How much to share with third parties? user privacy concerns and website dilemmas. *MIS Quarterly* 42(1):143–164.

Gu Y, Wenzel T (2014) Strategic obfuscation and consumer protection policy. *The Journal of Industrial Economics* 62(4):632–660.

Gupta R, Bagchi A, Sarkar S (2007) Improving linkage of web pages. *INFORMS Journal on Computing* 19(1):127–136.

Harsanyi JC (1968) Games with incomplete information played by “bayesian” players part ii. bayesian equilibrium points. *Management Science* 14(5):320–334.

Hölmstrom B (1979) Moral hazard and observability. *The Bell journal of economics* 74–91.

Jenson MC, Meckling WH (1976) Theory of the firm: managerial behavior, agency costs and ownership structure. *Journal of financial economics* 3(4):305–360.

Khatibloo F, Liu S, Pilecki M, Flug M, Hartig K (2018) Right your privacy ship before it capsizes. *Forrester* URL `http://www.forrester.com/report/Right+Your+Privacy+Ship+Before+It+Capsizes/-/E-RES133381`.

Krishnamurthy B, Naryshkin K, Wills C (2011) Privacy leakage vs. protection measures: the growing disconnect. *Proceedings of the Web*, volume 2, 1–10.

Li Y (2012) Theories in online information privacy research: A critical review and an integrated framework. *Decision Support Systems* 54(1):471–481.

Libert T (2015) Exposing the hidden web: An analysis of third-party http requests on 1 million websites. *arXiv preprint arXiv:1511.00619* .

Mare RD (1980) Social background and school continuation decisions. *Journal of the american statistical association* 75(370):295–305.

Pavlou PA (2011) State of the information privacy literature: Where are we now and where should we go? *MIS Quarterly* 977–988.

PewResearch (2019) Americans and privacy: Concerned, confused and feeling lack of control over their personal information. `https://www.pewresearch.org/internet/2019/11/15/americans-attitudes-and-experiences-with-privacy-policies-and-laws`, accessed: 2021-11-24.

Quintin C (2015) Healthcare. gov sends personal data to dozens of tracking websites. *Electronic Frontier Foundation* .

Roesner F, Kohno T, Wetherall D (2012) Detecting and defending against third-party tracking on the web. *Presented as part of the 9th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 12)*, 155–168.

Salesforce (2019) State of the connected customer. `https://www.salesforce.com/news/stories/state-of-the-connected-customer-report-outlines-changing-standards-for-customer-engagement/`, accessed: 2021-11-20.

Smith HJ, Dinev T, Xu H (2011) Information privacy research: an interdisciplinary review. *MIS Quarterly* 35(4):989–1016.

TechTimes (2020) Safe browsing made easy: Benefits of using web of trust. `https://www.techtimes.com/articles/248191/20200319/safe-browsing-made-easy-benefits-of-using-web-of-trust`, accessed: 2021-11-24.

Vincent J (2021) California bans 'dark patterns' that trick users into giving away their personal data. *The Verge,* *`https://www.theverge.com/2021/3/16/22333506/california-bans-dark-patterns-opt-out-selling-data`* .

Wilson CM (2010) Ordered search and equilibrium obfuscation. *International Journal of Industrial Organization* 28(5):496–506.

Yang Y, Liu H, Cai Y (2013) Discovery of online shopping patterns across websites. *INFORMS Journal on Computing* 25(1):161–176.

Yu Z, Macbeth S, Modi K, Pujol JM (2016) Tracking the trackers. *Proceedings of the 25th International Conference on World Wide Web*, 121–132.