Reputation and the Provision of Data Security

Manos Perdikakis*

August 22, 2024

Abstract

In a two-period model, a monopolist chooses unobserved data-security investments. Consumers pay no access fee, but must share their personal data and suffer when data breaches occur. The firm wants to earn a reputation for protecting users' data, to maintain high activity in period two. I analyse two regimes of endogenous data-sharing, differing as to whether the firm or the consumers have ex-post control over it. Starting at the firm-control equilibrium, the social planner can improve total consumer surplus by ex-ante imposing lower amounts of data-collection for both high- and low-reputation firms in the second period. On the other hand, compared to the ex-post consumer optimum, committing to less data-sharing following a breach induces higher security; the ex-ante optimal levels of data trade off the direct benefit of higher security against the cost of reduced learning about the level of cyber-risk. I discuss how these results relate to GDPR-type regulation regarding consumer consent, and also examine penalties and minimum security standards. Total consumer surplus is maximized by giving consumers control over data sharing and using penalties to discipline the firm's investment incentives.

*manos.perdikakis@economics.ox.ac.uk. Department of Economics, University of Oxford, and Jesus College. I am grateful to my supervisor, Margaret Meyer, for her guidance on this project and must thank Greg Taylor, Alexei Parakhonyak, as well as Yassine Lefouili, Giulio Gottardo and Andrew Rhodes, for very helpful discussions. This paper has also benefited from many helpful discussions with speakers of the Nuffield Economic Theory Seminar series, the Oxford IO Reading Group, as well as members of the department at the Toulouse School of Economics. I thank my discussants at the 15th Paris conference on Digital Economics, the 12th Oligo Workshop, and CRESSE 2024. I gratefully acknowledge funding from the Department of Economics at the University of Oxford, as well as the AG Leventis Foundation.

1 Introduction

Following the surfacing of major data breaches, most notably the Cambridge Analytica scandal, suspicion has arisen regarding the extent to which firms that handle personal-data respect their users' privacy. In Mark Zuckerberg's own words, the Cambridge Analytica scandal represented a "breach of trust" between Facebook and its users[1]. In the Senate hearing that followed the scandal[2], he pointed at the crucial importance of "trust" for Facebook's business model, which depends on maintaining *long-term relationships* with users who share their most personal information with the platform[3].

Trust is necessary because it is difficult for firms to credibly signal that they adopt good data-protection practices. Even when firms try to be fully transparent with their privacy policies, users often do not read them thoroughly, due to the texts being significantly long or convoluted or making extensive use of legal terminology. Or users may treat them as cheap talk, i.e. non-binding statements. Even in the presence of stated privacy policies, firms seem to have significant ex post *discretion* on how to implement those policies and as demonstrated in Mark Zuckerberg's Senate hearing following the Cambridge Analytica scandal, it may be hard to verify the extent to which breaches occur due to firms' poor security practices. This means we can plausibly think of firms' actions to provide high data-protection as both *unobservable* and *non-contractible*.

This discussion implies that *reputations* for good data-security could play a big role in the interaction between long-lived firms and privacy-concerned consumers. I find this consistent with the observation that firms seem concerned with convincing users that they value their privacy[4]. For another example, see Apple's recent campaign from 2023: "Privacy. That's Apple". [5]

[1]https://www.forbes.com/sites/kathleenchaykowski/2018/03/21/mark-zuckerberg-addresses-breach-of-trust-in-facebook-user-data-crisis/#3cc9c33d3e36, 2018

[2]https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/

[3]Senator Gary Peters' quote from the Senate hearing conveys the same message: "Can I believe...who has access to this information about me? So, I think it's safe to say, very simply, that Facebook is losing the trust of an awful lot of Americans as a result of this incident".

[4]Using Facebook as an example, its business website mentions: "..we take data protection and privacy very seriously and are committed to complying to data protection legislation..", while Mark Zuckerberg recently outlined his "Privacy-focused vision for Social Networking".

[5]https://www.apple.com/uk/newsroom/2023/01/apple-builds-on-privacy-commitment-by-unveiling-new-efforts-on-data-privacy-day/

This motivates me to investigate whether firms' concerns to maintain users' trust provide them with sufficient incentives to adopt data-protection practices. The concern is particularly acute because many digital service providers are mostly monetizing either consumer attention via ads, or more broadly consumers' data, offering their services for free, and may thus not have sufficient incentives to offer high quality services to their consumers. In such a world of reputational dynamics, I aim to understand the welfare impact policies that aim to give consumers more control over data-sharing decisions and curb the monetization of personal data.

In my benchmark model, I examine a two-period interaction between a monopolist service provider and consumers, who do not pay a monetary price to access a firm's service, rather must share their personal data with the firm. The firm monetizes this information and chooses its level of *unobservable* data-security investment in order to avoid breaches of its database. I will be considering consumers who value their privacy, and with each data breach suffer disutility, which increases in the amount of data that they must share with the firm in exchange for using the service. In the first section, this amount of data-collection will be treated as exogenously determined.

In addition to unobservable security investments, I use a model with consumer *uncertainty* about firm characteristics that determine the probability of a data-breach, to capture the fact that consumers are uncertain about the riskiness of sharing their personal data with a given firm. They will thus rely on the occurrence or not of data-breaches to *learn* about the risk of sharing data with the firm. After updating their beliefs, they make their activity choices again in the second period.

In a two-period model, firms will be motivated to invest in the first period in order to avoid public data breaches and the resulting harms to their *reputation* as adopters of good data-protection practices. In terms of modelling, I will be using a model of in which firms may be Commitment or Normal types, and the Commitment type is non-strategic and always provides high level of data-security. A lack of data-breaches is good news, and makes consumers update their belief upwards about the probability they are facing a Commitment type; I will refer to this belief as the firm's *reputation*. Absent regulation, investment incentives are purely implicit, motivated by consumer retention, and the Normal type makes no security investment in the last period.

The equilibrium derivation is followed by the main welfare and policy analysis of this paper. I extend the model to study a setting in which the level of data-sharing is endoge-

nous and consider two different modes of endogenous data-sharing: the *consumer-control* regime in which consumers make **ex-post** optimal data-sharing decisions, given a firm's current reputation, and a *firm-control* regime, in which the firm chooses the amount of data consumers are required to share with the firm, in order to access the service. The firm chooses the requirement in order to ex-post maximize its profit, taking its current reputation into account.

I'm motivated to study these regimes of data-collection control because I believe they map accurately to how cookies-related data-collection took place before and after the introduction of the EU GDPR "opt-in" regulation. Under the latter, which is also the subject of the empirical study[6] by Aridor et al. [2023], firms must ask for consumers to explicitly opt into data collection and consent is required for *each purpose of data-processing individually.* Prior to this regulation, firms did not need to offer opt-out options and it is plausible to assume that they chose their cookie-collection in a profit-maximizing manner. The direction of the GDPR towards giving consumers more choice in data-provision, also motivates the related work by Markovich and Yehezkel [2021] to compare between firm- and consumer-control regimes of data-collection, and in that paper, the two regimes are modelled in the same way as in mine[7].

Starting at the equilibrium of either of those regimes, I ask: does ex-ante commitment to different *history-dependent* data-sharing levels allow a planner to increase expected consumer surplus relative to equilibrium? The planner observes no informational advantage over consumers but can commit to different data-sharing levels depending on the firm's reputation, i.e. whether it suffered a data-breach or not. I will be asking this question around the equilibrium of each of the two regimes of ex-post data control, similar to the approach taken by Lefouili et al. who also examine the welfare impact of caps on data-monetization by a firm.

Changes in the levels of data-sharing in the second period have multiple effects on consumer surplus (CS) in this model: the direct, on the utility of active users in the second period, and the indirect, via changing equilibrium investment incentives. In turn, equilibrium investment affects consumer surplus via first-period disutility from breaches, and via increasing the relative frequency with which a Normal (low-security) type will have a high reputation in period two. Conditional on facing a Normal type, consumers face ex-

[6] The authors find a reduction in total cookies by 12.5% caused by this regulation.

[7] Although there is no cyber-security or learning in their paper and the emphasis is on consumer heterogeneity with respect to privacy preferences and on data-externalities.

post regret in equilibrium following no-breach; they share too much data (in either regime) and too many users are active relative to a perfect information setting, because they entertain the possibility of facing a Commitment type. I call this the "signal-jamming" effect of higher investment and it is always *negative*: high first-period secuirty impedes learning and reduces second-period CS.

Starting from the data-collection values of the equilibrium under firm-control, changes in the levels of second-period data sharing have no first-order impact on investment incentives; data-sharing affects investment incentives via changing profits in each of the second period states, and at the ex-post profit maximizing levels of data sharing, profit is insensitive to changes in them. This means that the only first-order impact is the *direct* one on second-period CS. A profit maximizing firm whose revenue per consumer increases with data-sharing will always ask for so much data that CS is decreasing at the margin, so that the direct effect of limiting data-sharing is positive. A CS-maximizing planner who faces a regime of firm-control can therefore set small caps on the levels of data-collection in period 2, on *both* high- and low-reputation firms and achieve an increase in *total* CS.

On the other hand, a planner that faces a regime of consumer-control deals with a different situation; in that case, the direct effects on expected second-period CS are zero because data-sharing in period two is chosen optimally by consumers. Therefore, consumers can benefit in the second-period by changes in data-sharing that induce *less* equilibrium security, so that there is more accurate learning about the environment, i.e. less signal-jamming. However such a reduction will come at the expense of first-period security. This is the fundamental policy trade-off that emerges in this model, because of the dual role investment has. It both affects real outcomes, but also impedes learning about the firm's type. At this equilibrium, security investment may be too high or too low relative to the consumer-optimal level. Data-caps for high- and low- reputation firms have effects of opposite direction on equilibrium investment, so that the nature of intervention is different according to the firm's reputation.

~~I then extend the benchmark model to a duopoly; I find that with linear revenue, equilibrium investment of each firm is always lower relative to monopoly; this is a simple consequence of the fact that under linear revenue in market share, the presence of a competitor will reduce the marginal benefit to achieving high reputation in the second period. As the previous analysis suggests, this does not necessarily imply lower consumer surplus, since it will imply faster learning about firms' types and less ex-post regret in the~~

second period. In Appendix C, I introduce endogenous data-sharing and examine how data caps can affect consumer welfare in a duopoly, where firms simultaneously choose their required levels of data-sharing to attract consumers. I use mostly numerical simulations to find that data caps can consistently increase consumer surplus relative to the firm-control optimum, despite the fact that competition drives firms' equilibrium data extraction down.

Finally, I use the benchmark model to identify how reputational incentives interact with common policies. I first examine the impact of two policies on equilibrium investment: penalties on firms that get breached and the specification of minimum security standards. Both of those policy levers act as *commitment devices* in my model. High level of a minimum standard means that even if a firm has low reputation following a data breach, consumers understand that it will use security at least equal to the mandatory minimum in the second period. But this decreases the harm to the firm from having low reputation, thus erodes the implicit incentive to achieve high reputation in the first period. Unless the planner can specify a sufficiently high level for minimum security standards, adopting such a policy will increase second-period investment but *decrease* first period investment in equilibrium.

To further motivate the model, it is worth it looking at some literature which suggests that firms might indeed suffer financial damage following a data breach. Focusing on public corporations in the US, Kamiya et al. [2021] find significant negative abnormal returns only when cyber attacks induce the loss of personal data; the abnormal returns of firms that do experience negative returns are almost 500 USD million per attack (1 percent of value). Closely related to my model of reputation incentives, the authors argue that in a full-information world where there is no learning about the firm or the environment after a successful cyber attack is disclosed, the firm's loss of value should only reflect *out of pocket fees* (e.g. penalties, legal fees, etc.). Using data on disclosed breaches from 2005 to 2017, they estimate that cyber attacks have substantial additional **reputation** costs on top of those due to expected legal action and penalties. Reputation in their setting, and in mine, is synonymous with the firm-specific distribution of losses due to cyber attacks that the customers perceive. Thus, their paper provides valuable empirical justification both for the learning component of my model and the existence of firm incentives to avoid data breaches.

Even though I have drawn motivation from the Cambridge Analytica case, the concerns described above are not restricted to social media. The example that motivates the

analysis of closely related work in Jullien et al. [2020] can be effectively used to provide motivation for my work too; the authors recall an incident in which the Times website, due to *insufficient diligence* in screening third-parties that were allowed to post ads on the newspaper's website, exposed its users to digitally harmful material.

In the next section, I discuss related literature. In Section 3, I present the benchmark monopoly provider model, with exogenously determined levels of data sharing. The main body of policy and welfare analysis is in section 4, in which I introduce the two regimes of endogenous data collection, and discuss the ability of a planner to improve consumer surplus by pre-committing to levels of second-period data collection that depend on a firm's posterior reputation. Section 5 introduces the extension of the baseline model to a duopoly, and Section 6 shortly analyses two simple policies in the duopoly context with exogenously determined data collection. The paper then concludes.

2 Related Literature

In this literature review, I find it more worthwhile to discuss few papers in greater length that are closest to mine, rather than attempt to list all papers in the large literature about the economics of privacy and cyber-security provision. There are excellent surveys both very recent by Goldfarb and Tucker [2023], as well as slightly older by Acquisti et al. [2016]. The former also covers the recent empirical work, both on the economic impact of GDPR and on measuring privacy concerns. The latter deals in depth with the theory literature on the economics of privacy.

The paper closest to mine is probably Jullien et al. [2020]. Their model uses a signal-jamming, two-period model of belief formation and in their paper too, firms take unobserved actions, in the form of screening the third-parties they share consumer data with. In their model, as in mine, equilibrium incentives are based on the prospect of consumer retention. However, consumers do not update their beliefs about firm attitudes towards privacy, rather about their own *vulnerability* in the event of a data-breach. Their single-website model is similar to my model of monopoly with fixed data terms, but the focus of their paper is multi-homing competition between websites.

They study this mode of competition for consumers in order to focus on (a) website competition in the *advertising market* and (b) on a novel "public good" problem between websites: as long as consumer vulnerability is positively correlated across websites, a lack

of precaution by any one of them means that the consumer is more likely to value using any of the other websites in the next period less. The public-goods aspect means that in the perfect correlation case, a "zero-protection" equilibrium always exists. In contrast to their analysis, I distinguish between data-collection and data-protection, and I focus on *history-dependent* caps on data-sharing and analyse different regimes of endogenous data-collection.

Also related is the paper by Lefouili et al., which is again motivated by regulation that requires informed consent for data processing. The authors examine how caps on data-monetization[8] will affect firm incentives for *observable* investment on quality and examine the trade-off between higher quality and more privacy. An imposed ceiling on data-monetization[9] will increase the amount of data shared with the firm by privacy-conscious consumers, which in turn increases firm incentives for quality investment to the extent that higher quality will attract more data-sharing consumers.

The papers by Markovich and Yehezkel [2021] and Dosis and Sand-Zantman [2023] draw welfare comparisons between consumer-and firm-control of data-collection decisions, also motivated by GDPR-style regulation which gives consumers greater control. Neither studies cyber-security. In Markovich and Yehezkel [2021], which regime is optimal depends on the magnitude of data-externalities: large and positive externalities implies that the firm regime dominates, given atomistic consumers would under-supply. In Dosis and Sand-Zantman [2023], the firm has less incentive to invest in data-processing under consumer-control, so that regime is optimal only if the gains from more privacy dominate the loss from lower data-utilization.

The following papers, de Cornière and Taylor [2021], Ahnert et al. [2022], and Fainmesser et al. [2023], all have as their main focus the impact of the firms' business model on equilibrium incentives for privacy provision or cyber-security investment. They differ from mine in that they model the incentives of cyber-attackers and the frequency of data-breach *attempts* is endogenous in their papers. They all use static models and do not consider reputation-based incentives for security.

In the first one, de Cornière and Taylor [2021], the authors study the interaction between the firms' business models in duopoly and equilibrium levels of cybersecurity. They

[8]The authors distinguish between data shared with the firm and the amount of data that the firm monetizes. The two distinct values both enter consumers' utility functions.

[9]Similar to my work, the authors employ a local analysis around the profit-maximizing level of data-monetization.

do this in a static setting with observable investments. The introduction of strategic hackers introduces a *negative* network externality between users of each firm, since a large user base attracts more data-stealing attempts. Their discussion focuses on comparing equilibrium investments between ad-funded duopolists and product funded ones that charge (endogenous) prices. The fact that they use observable security decisions, leads to different findings than mine regarding the efficiency of investment provision in monopoly[10] and duopoly, in the comparable advertising regime, compared with the extension of my model to a duopoly setting.

Ahnert et al. [2022] is a model of security provision and fee choice by financial intermediaries. Unlike de Cornière and Taylor [2021], they study a single business model of the firm that interacts with consumers, but they study different modes of operation by the hackers, who can either choose to ask the firm for ransom or engage in conventional attacks and attempt to steal users' data. Attackers first choose their mode of operation which the firm observes, then the firm chooses fee and security level (which the users may or may not observe, they deal with both cases) and then attack commences. Both papers study the optimal design of liability as well as minimum security standards.

Fainmesser et al. [2023] also models attackers' side in detail. Their innovation is dealing with both data-storage and data-sharing choices of the firm and they analyze those, both for ad-funded and transaction-funded firms. They take a firm's business model as given and find the optimal data collection and data security levels. Firms that are more data-driven, set both higher levels of data collection and protection. This complementarity arises because higher collection attracts more attackers and thus raises the marginal benefit of protection.

To the best of my knowledge, and according to the survey by Goldfarb and Tucker [2023], and there is no empirical work documenting the impact of GDPR on cyber-security investments and equilibrium frequency of data-breaches[11].

Koutroumpis et al. [2022] examine the link between hiring of cyber-security specialists by firms and stronger data-protection laws and enforcement in the UK, and find significant positive impact of the new policy on cyber-security hiring expenditure, using Burning Glass job ads data. They focus specifically on data-breaches as a subset of cyber-attacks, since

[10] In their "monopoly" example, they assume full market coverage, hence there exist no incentives for the provision of security.

[11] As Garrett Johnson notes, "we have seen more research on the *unintended* consequences of the GDPR, rather than the *intended*".

those both (a) involve a loss of personal information and (b) are often harmless to the victim firm in terms of operations disruption. These two elements lead to potential *under-investment*. This is also the motivating application I will make use of for this paper, i.e. I will not be thinking about *ransomware attacks*, because they are both directly harmful to the firms that have to pay ransom to restore some part of their digital operations, and could plausibly be costless to consumers, if the firms pay the ransom and attackers are 'noble' in the sense that they don't sell data even after receiving a ransom payment.

3 A model of data-breaches

There is a continuum of consumers with mass one, uniformly distributed over a line on $[0, 1]$. They interact with a single firm over $T = 2$ periods. The firm provides a digital service and wishes to attract users; Registration of users lasts 1 period, while firm and users live for 2 periods. Users make their participation decisions at the beginning of each period. The firm charges users *no registration or usage fees*, but users must share their personal data with the firm in order to use the service[12]. Firms have some ex post discretion on how much to *invest* in the protection of their users' data. This investment can be thought of as effort that firms exert to better screen third parties that get access to consumer data, or as actual investment in cyber-security to deter data breaches. This variable will be denoted by e and I refer to it as the effort/investment/security level. Crucially, I assume that this effort is *unobserved* by the consumers and *non-contractible*. I find this assumption reasonable; even if a data breach is made publicly known, it could be quite costly, if at all feasible, to prove that it was due to lack of due diligence by the firm.

Towards attracting privacy-concerned potential users, the firm faces potential gains from maintaining a *reputation* of caring about users' privacy. I reputation by introducing incomplete information about firms' *types*; a firm can have type N or C, which stand for Normal and Commitment type, respectively. This type is **privately known** to the firm. The Commitment type is non-strategic and always chooses the same action. In particular, a Commitment type is the "good" type and always chooses the highest level of effort, $e = 1$. Similar to Benabou and Laroque [1992], we can interpret this modeling device as representing either real uncertainty about firm attitudes, or, perhaps more realistically,

[12]Or we could think of this as data being generated by their activity on a website/app, which the firm can subsequently monetize.

uncertainty about the firm's payoffs: a Commitment type can then be thought of as a firm which incurs sufficiently high monetary loss from a data-breach in expectation, that it finds optimal to play $e_t = 1$ in every period. In this paper, we will be concerned with equilibrium incentives of Normal types.

As I will describe in more detail in the next subsection, consumers care about the level of effort the provider exerts, because this effort level determines the probability with which they experience a breach of their personal data and suffer privacy disutility. I define the "outcome" binary random variable s_t, which can take values $\{b, n\}$, standing for *breach* and *no breach*. The value of this random variable becomes publicly known at the end of each period and $P(s_t = b|e_t)$ is the probability it takes value *breach*, given the firm's effort level[13] e_t of period $t \in \{1, 2\}$. I use the following specification for the conditional pmf of s:

$$P(b|e) = \zeta + (1 - \zeta)(1 - e) \tag{1}$$

And taking expectation across types for given reputation μ:

$$p(\mu, e) := E_\mu[P(s = b|e)] = \zeta + (1 - \mu)(1 - \zeta)(1 - e) \tag{2}$$

The interpretation of the above pmf is simple: in each period, there is a probability $\zeta \in [0, 1]$, that a negative breach "shock" will arrive regardless of the firm's effort choice. As the above specification suggests, breach probabilities in period 2 are *independent* of s_1 and e_1. I will refer to the firm's *reputation* in a given period, as the probability with which users believe that the firm's type is C in that period. The firm has prior reputation $P(C) = \mu_1$, which is common knowledge. After observing $s_1 \in \{n, b\}$ at the end of $t = 1$, fully rational users update their beliefs the firm's type using Bayes' Rule. The posterior reputation, $\mu_2 \in \{\mu_n, \mu_b\}$, depends on the prior and also on the effort level that users believe the Normal-type exerts in the first period, $\tilde{e_1}$. I assume all users share the same conjectures, and everybody observes the realization of s_1, so that there is a single posterior reputation for the firm at the end of $t = 1$.

For any $\zeta < 1$, the posterior reputation a firm achieves following a $s_1 = n$ realization in period one is:

$$\mu_n(\tilde{e_1}) := P(C|s_1 = n, \tilde{e_1}) = \frac{\mu_1}{\mu_1 + (1 - \mu_1)\tilde{e_1}} \tag{3}$$

[13]The only channel via which the firm's type influences the probability of a breach is via investment, e_t. I do not need to condition the probability on the firm's type, when conditioning for e.

The above posterior is not well-defined for $\zeta = 1$, since the probability of a non-breach outcome becomes zero in that case. Posterior reputation following a good outcome takes values in $[\mu_1, 1]$. For any $\tilde{e_1} < 1$, a lack of breach is relatively more likely when facing C type and $\mu_n > \mu$. Expression (3) further reveals that posterior reputation is decreasing in the effort conjecture, $\tilde{e_1}$. This is very intuitive; if users anticipate that a Normal type firm exerts a lot of effort to prevent breaches from occurring, a lack of breach becomes less informative about the firm's type, less indicative of a Commitment type and thus lower positive impact on the firm's prior reputation[14]. When $\tilde{e_1} = 0$, then all good news indicate a Commitment type with certainty, so posterior reputation following $s_1 = n$ is equal to one. In contrast, when $\tilde{e_1} = 1$, the Normal type firm perfectly replicates the Commitment type's behaviour in period 1, hence posterior reputations are not updated and $\mu_n = \mu_1$. Note that μ_n does not depend on ζ since the ratio of probabilities with which each type achieves a "no-breach" realization is constant[15] with respect to ζ.

Similarly, the posterior reputation for a firm that has a "breach" outcome in period 1 is:

$$\mu_b(\tilde{e}_1) := P(C|s_1 = b, \tilde{e}_1) = \frac{\zeta\mu_1}{\zeta\mu_1 + [(1-\zeta)(1-\tilde{e_1}) + \zeta](1-\mu_1)} \tag{4}$$

which now is always *smaller* than μ and is *increasing* in ζ. This is intuitive; for $\zeta = 0$, we are in a perfectly revealing bad news setting and a breach realization lets consumers know that they are facing an N type with certainty. Higher ζ allows consumers to entertain the possibility that the breach was a result of a negative shock. This posterior is *increasing* in the consumers' effort conjecture, since a bad result is more likely to be the outcome of a negative shock rather than firm negligence (low effort).

We can think of ζ as inversely related to the quality of public infrastructure and support given to firms to protect against cyber warfare. For instance, as the level of support that firms receive in terms of information provision regarding state-of-the-art cyber attacks. Similarly, we can think of ζ as the probability in each period that firms are attacked using highly sophisticated hacking methods that they could not have protected themselves against, or simply as the minimum probability that firms are exposed because of human error in their processes (e.g. an employer losing their work laptop). Throughout

[14]This is in contrast with other models of effort provision: For instance, in models using the setup of Holmström [1999], posterior reputation is not a function of effort conjectures, hence the resulting first-order condition is linear in the simplest model.

[15]That is because ζ is a shock that results in a breach regardless of investment, which is the only difference between types.

this paper, I will use the b subscript to denote variables for the period-2 state after a breach has occurred and the subscript n in the same manner. For example, $p_b(\mu, e)$ or just p_b will be the expected probability of a breach given reputation μ_b. I will also be writing μ_n, μ_b instead of being explicit about their dependence on $\tilde{e}_1$.

3.1 Consumers

I now turn to users' payoffs and participation decisions. As mentioned already, users make their participation decisions at the beginning of each period, meaning that users choose between using the firm's service or staying idle. Each user is characterized by a type θ, which is the value of their outside option and follows distribution F. Active users derive positive utility $v(d)$ from using the service, where $d \in [0, d^{max}]$ is the amount of *data* that they share with the firm. However, users also suffer disutility[16] $\ell(d)$ in the event of a *data-breach* and $\ell'(d) > 0$ so that users suffer more from a breach when they have shared more data. There is no heterogeneity in privacy preferences. I think of d as the data input required by consumers to use the service. It can potentially differ across periods, in which case I will use a time-subscript, and in this section I will treat it as *exogenously* given. Expected utility given probability of a breach, p, is then:

$$u(d, p) = v(d) - p\ell(d) \tag{5}$$

I will be assuming that u is *quasi-concave* in d, for every $p \in [0, 1]$. Thus, consumers can potentially benefit from sharing at least some data with the firm. Furthermore, a user that has suffered a breach in the first period will incur additional loss of $\ell(d_2)$ if their data is breached again in the second period. The disutility of an active user that experiences a breach in the second period is independent[17] of both first-period activity and s_1. Last, but not least, I abstract from network effects and informational externalities by assuming that the utility users derive from using the service is independent of other users' participation decisions (both past and present). If informational externalities in the spirit of Acemoglu

[16] Lin [2022] attempts to disentangle between "taste" for privacy and instrumental preferences, i.e. preferences stemming from anticipated surplus loss in the absence of privacy. Using a lab experiment, the paper finds that consumers do have both *intrinsic* and instrumental preferences for privacy. In my model, I will not take a stance on whether consumers' privacy preferences are intrinsic or instrumental.

[17] This means, for example, that an agent that uses the service in both periods does not incur higher privacy cost in the event of a second-period breach than a user who just uses it in $t = 2$. Users suffer only because their current period data is exposed.

et al (2022) were present, then the outside option would be weakly negative for some users with low θ and decreasing in the mass of active users.

3.2 Equilibrium

Given the above, the mass of active users in a given period is $F(u(d,p))$. The firm earns revenue $r(d)$ per active user, which is net of the constant marginal cost of servicing an additional consumer and increasing in the amount of data collected per user, i.e. $r'(d) > 0$. We define the firm's revenue as $\Pi(d,p) := r(d)F(u(d,p))$. In the first period, the Normal type chooses e_1 to maximize expected profits across both periods net of investment cost:

$$\begin{aligned} \mathrm{E}\Pi(d, e_1; \tilde{e_1}) = \Pi(d,p) - C(e_1) \\ + P(b|e_1)\Pi(d,p_b) + (1 - P(b|e_1))\Pi(d,p_n)) \end{aligned} \tag{6}$$

where $P(b|e_1) = \zeta + (1-\zeta)(1-e_1)$ is the actual probability of a breach given a Normal type, whereas p is the consumers' expectation defined in (2). The cost function is increasing and convex in e with $\lim_{e\to 0} C(e) = \lim_{e\to 0} C'(e) = 0$, e.g. $C(e) = \frac{1}{2}ce^2$, and the Normal-type firm chooses effort to maximize the above profit function, taking consumers' conjecture $\tilde{e_1}$ as given. The firm bears no direct loss in the event of a breach, thus investment in security is only motivated by concerns to attract users in period 2, implying that $e_2 = 0$.

Demand in the first period and posterior reputations μ_n, μ_b only depend on consumers' *conjecture*, $\tilde{e_1}$, and are not directly influenced by the firm, even though that conjecture will have to be correct in equilibrium. The first-order condition that must be satisfied at an interior solution is:

$$(1-\zeta)\Big(\Pi(d,p_n) - \Pi(d,p_b)\Big) = C'(e_1) \tag{7}$$

where I am using the shorthand notation $p_n = p(\mu_n(\tilde{e_1}), 0)$ and $p_b = p(\mu_b(\tilde{e_1}), 0)$. Equation (7) defines the monopolist's optimal[18] effort provision, as a *best-response* to consumers' investment conjecture, $\tilde{e_1}$. Greater difference between revenue in the two potential outcomes induces higher investment provision. To turn (7) into an equilibrium defining equation, I must impose the equilibrium condition that conjectures are correct i.e. $\tilde{e_1} = e_1$. Since μ_n is *decreasing* in $\tilde{e_1}$ and[19] μ_b is *increasing* in it, the *equilibrium* marginal benefit curve is downward sloping and we obtain equilibrium existence and uniqueness.

[18]The solution to this first-order condition is always the global maximizer, since the marginal benefit is independent of the actual investment.

[19]For $\zeta > 0$. Otherwise, $\mu_b = 0$ for all $\tilde{e_1}$.

Proposition 1 *A unique Perfect Bayesian Equilibrium of the monopoly game exists for all parameter values. Type C plays $e = 1$ in both periods; type N plays $e_1 = e^*$ and $e_2 = 0$. In equilibrium, users' conjectures are correct, i.e. $\tilde{e_1} = e^*$ and $\tilde{e_2} = 0$. First-period choices by the Normal type maximize expected profit* (6) *given those conjectures.*

- *If $\zeta = 0$, e^* is given by the unique positive solution to the equilibrium first-order-condition, if the latter is weakly lower than 1. Otherwise, it is given by the corner solution $e^* = 1$, and we have a pooling*[20] *equilibrium.*
- *If $1 > \zeta > 0$, e^* is given by the unique solution to the equilibrium first-order-condition* (7) *in $[0, 1]$ and always lies strictly between $(0, 1)$.*
- *If $\zeta = 1$, positive investment cannot be supported in equilibrium, $e^* = 0$.*

A more careful proof is in the Appendix, alongside proofs for the following comparative statics results:

Lemma 1 *The equilibrium effort level e^*:*

1. *Is **increasing** in d, for $\theta \sim U$.*
2. *Is **decreasing** in ζ, for ζ sufficiently close to 0. For $\theta \sim U$, it is decreasing in ζ for all $\zeta \in [0, 1)$.*
3. *For for two functions that satisfy $r_1(d) > r_0(d)$, $e_1^* \geq e_0^*$.*

To interpret the above comparative statics results, arguments under fixed conjectures suffice. In other words, we can consider changes in the marginal benefit of investment without considering the feedback from consumers' conjecture $\tilde{e_1}$ that changes in equilibrium. Equilibrium is found at the intersection of a downward sloping curve of the firm's best-response to consumers' conjecture $\tilde{e_1}$ with the $e_1 = \tilde{e_1}$ line. This implies that the total effect of a change in parameters on equilibrium investment will be of the *same direction* as if beliefs were fixed, but also of *lower magnitude*.

The first result obtains because $r'(d) > 0$, but also $u_{d\mu} > 0$, i.e. the mass of active users is more sensitive to reputation (i.e. the probability of a breach) at higher levels of d. These effects suggest $\frac{\partial^2 \Pi}{\partial d \partial \mu} > 0$, which leads to the reported comparative statics result.

[20] In the sense that type N perfectly imitates type C in period 1 – but their period 2 behaviour is still different.

The exogenous shock probability, ζ, has multiple effects on the left-hand side of (7), i.e. the marginal benefit of investment (for fixed $\tilde{e_1}$). First, a *negative* direct effect is that investment is less effective in reducing the probability of a breach. Second, an increase in ζ increases μ_b and leaves μ_n unchanged, which is also a negative indirect effect via posterior reputations. Third, there is an additional negative direct effect because an increase in ζ also affects breach probabilities in the *second* period. The Normal type is always breached in the second period, whereas the Commitment type is always breached with probability ζ. Thus, an increase in ζ increases the perceived probability of a breach by more, in states where consumers believe to be facing a Commitment type with higher probability. In other words, $\frac{\partial^2 p}{\partial \zeta \partial \mu} > 0$.

It is interesting to think about the interpretation of ζ as a parameter that a regulator can affect. Apart from the direct gain of reducing ζ, a regulator would also indirectly by increasing the effort induced by Normal type firms.

It is non-trivial to establish whether a firm with a higher prior μ will exert higher effort in equilibrium, since both posterior reputations present in the net gain term are increasing in the prior, holding conjectures fixed. In the case of perfect bad news where $\zeta = 0$ and μ_b becomes zero, equilibrium effort would be increasing in the prior because the expected gain from not getting breached would be higher for every conjecture level held by consumers. By continuity of e^* in ζ in a neighbourhood of $\zeta = 0$, this result will carry-through for low values of this parameter.

3.3 Disclosure of breaches

We can extend the model by introducing imperfect disclosure of data-breaches. We do so in a reduced form way, via the parameter $q \in [0, 1]$, which is the probability that a breach becomes public information after it has occurred. For simplicity, assume that the same q applies for breaches that occur to both Normal and Commitment firms. For $q < 1$, consumers update their beliefs at the end of period 1 based on whether or not they *observed* a breach. The probability with which this happens is $\kappa^b(e, \zeta, q) := [\zeta + (1 - \zeta)(1 - e)]q$, increasing in q and ζ and decreasing in e. Since the likelihood ratio $\kappa^b(e, \zeta, q)/\kappa^b(1, \zeta, q)$ does not depend on q, the posterior μ_b is also unaffected[21] by q. On the other hand,

[21]This relies on the assumption that the Commitment type's breaches are also only revealed with probability q. If they were always *decreasing* in q, then the ratio would become *increasing* in q and μ_b would be *decreasing* in it; this is intuitive under this alternative assumption: at higher q, an observed breach

the posterior μ_n is always *increasing* in q, since at higher disclosure rates, no disclosure becomes more informative about an actual lack of breach. The level of q has an additional *direct*, positive effect on investment incentives, and we obtain the following:

Lemma 2 *The equilibrium level of investment e^* is increasing in the probability of disclosure, q.*

3.4 Discussion of assumptions

Before moving on to the policy analysis and model extension, I discuss some assumptions on which the equilibrium derivation and subsequent analysis does not depend on qualitatively.

1. User heterogeneity in data preferences can be accommodated; what does matter is firm revenue in each period is decreasing in the probability of a breach that users perceive. New working paper by Lin et al finds heterogeneity in sharing data both across consumers (what I just discuss here) *and* across websites, which could suggest a role for website reputation to affect sharing decisions.

2. The quasi-concavity assumption on u can be guaranteed if $v''(d) < 0$ and $\ell''(d) > 0$. The latter is not obviously the most plausible assumption. It corresponds to an interpretation of higher levels of d as including more sensitive data whose leakage would be even more disliked by privacy-concerned consumers. Alternatively, we could think of d as the quantity of similarly sensitive data; in that case, $\ell''(d) > 0$ could be justified if targeting or identifying the consumer by malicious parties was increasingly easy or increasingly accurate with more data. A data breach gives such malicious parties access to consumers' data.

3. The firm need not be privately informed about its type. the model would work very similarly[22] as a pure "signal-jamming" model, in which some firms are type C and some are type N, but the firm itself also does **not** know its own type. The marginal benefit on investment to a firm that does not know its own type would have to be

becomes more likely to have originated from a Normal type.

[22] More complicated would be to introduce q as in the previous section. Under mutual uncertainty about the firm's type, discussing the disclosure of breaches would raise the question of whether the firm also learns of the breach itself. If it does, and consumers do not, then the firm would have private information in the second period. In a model of more than two periods, this would cause a qualitative divergence between the models with and without private information on the firm's side.

multiplied by $(1-\mu_1)$, since investment is only valuable if the firm is Normal type (in this case without private information, the types would be more clearly interpreted as high- and low-risk). This modelling assumption may be more appropriate if we are thinking as data-breach risks coming from, for example, zero-day vulnerabilities, the existence of which the firms are reasonably assumed to be unaware of. Finally, note that the firm's type does not need to be time invariant, just positively correlated across the two periods, for investment incentives to be supported in equilibrium.

4. Less importantly, the Commitment type could play $\hat{e} < 1$ and the equilibrium derivation follows identical arguments. Notice that with purely reputation-driven incentives, e^* cannot over-shoot $\hat{e}$ in equilibrium. In that case, a *breach* would be evidence of a Commitment type, hence no investment incentive in equilibrium.

5. The model can be extended to allow for a simple treatment of positive consumption externalities. Sufficiently high magnitude means the monopolist achieves full market coverage in period two regardless or μ_2, thus has no investment incentives. For modest magnitude, the analysis remains qualitatively the same.

Finally, the model presented above can be extended to accommodate different amounts of data collection in period 2, depending on the outcome of period 1, which we can refer to as d_n and d_b. When d_n and d_b are exogenously fixed parameters, equilibrium derivation follows the same arguments as above and each pair d_n, d_b induces a unique equilibrium. In the following section, I extend the model in order to endogenously determine the levels d_n and d_b.

4 Endogenous data collection

In this section, there are two objectives: First, to extend the monopoly model just presented and account for *endogenous* choices of the data variable; in particular I will be thinking about history-dependent choices, i.e. two separate values $\{d_n, d_b\}$ which refer to the two possible states of period 2. I will be focusing on two different regimes of *ex-post* control over data sharing. In the regime of *consumer control*, consumers can choose in every period the amount of data they want to share with the firm, if they participate at all. They can thus react to new information about the firm in the second period, by changing how much data they share with it to maximize their second-period expected utility.

Maintaining the initial assumptions of no data-sharing externalities between consumers, decisions under the consumer regime maximize ex-post consumer surplus. I further maintain the homogeneous data preferences assumption, so that there a unique level of data level that maximizes each consumer's surplus from using the firm's service.

In the regime of *firm control*, the firm chooses its profit-maximizing data requirement in each period and state; I use *ex-post* control, in the sense that the firm cannot commit in period 1 to how much data it will ask for consumers to share in period 2.

The second objective in this section is to understand, for each regime, whether (and how) a planner can raise consumer surplus relative to the "regulation-free" equilibrium. In particular, I will assume that the planner can impose a specific level of data to be shared by active users in each state of the second period, i.e. following either a breach or lack of one. Regarding the consumer regime, this is equivalent to asking whether consumers would collectively benefit from committing to different levels of d_n and d_b than those that ex-post maximize consumer surplus. Throughout this section, I will be focusing on the case of uniformly distributed consumer outside options. The mass of active consumers will be given by the indifferent type, θ_1, θ_n, or θ_b and I focus on the case such that these are always interior.

4.1 Equilibrium in the two regimes

For each regime, the timing of the game with endogenous data collection is almost identical to that of the previous section.

In the consumer regime, in each period, consumers choose both whether to be active users, and if so, the level of data to share with the firm, $d^C \in [0, d^{max}]$. In the second period, they do so after having formed posterior beliefs about the firm's type. I emphasize that this model preserves a feature of the previous analysis, namely that the consumer type is not interpreted as data-sensitivity, i.e. $u_{\theta d} = 0$. This means that consumers always agree on the optimal level d^C. For given probability of suffering a breach, active users share $d^C(p) := argmax_d\, u(d, p)$, which is uniquely defined if we assume $u(d, p)$ to be quasi-concave in d. In addition, the *negative* sign of $u_{d,p}$ implies that d^C is *decreasing*. The direct effect of p on expected utility is always negative, so that an increase in p decreases the mass of active consumers, too. Finally, a lower p increases firm revenue via both greater demand and greater revenue per consumer. In other words, if we define $\Pi^C(p) := \Pi(d^C(p), p) = r(d^C(p))D(d^C(p), p)$, we obtain the intuitive $d\Pi^C/dp < 0$.

In the firm regime, at the beginning of each period, the firm announces the level of data $d^F \in [0, d^{max}]$ that a consumer must share with the firm in order to use the service[23]. Importantly, we assume there are no data-adjustment costs between periods: thus, Normal and Commitment type firms have the same optimal d^F choices at every period and for every current reputation level, and there is no signalling[24] of the firm's type from the choice of d_1. We define $d^F(p) := argmax_d\, \Pi(d,p)$ and $\Pi^F(p) := \Pi(d^F(p),p)$, and using a similar envelope argument to the previous paragraph, we obtain that $d\Pi^F/dp < 0$.

For any p, d^F must be weakly larger than d^C: a firm decision cannot be profit maximizing if marginally increasing d would both increase demand and revenue-per-consumer. This means that in equilibrium of the game with firm-control, consumers would always rather that the firm asks for less data in each state. I will refer to this feature often, so it is useful to state it as a Lemma.

Lemma 3 *A firm that chooses d to maximize current-period profits will optimally choose a level d^F that satisfies $u_d(d^F, p) < 0$ or it will choose $d^F = d^{max}$. If, for any $p > 0$, Π and u are quasi-concave in d, it additionally holds that $d^C(p) \leq d^F(p)$, with strict inequality if $d^F(p) < d^{max}$.*

To generate the Figures that follow I will be using the functional forms:

$$u(d,p) = \alpha d - (p+1)d^2$$
$$r(d) = r_1 d$$

for any $\alpha, r_1 > 0$, the utility and revenue functions are quasi-concave in d, with the following maximizers:

$$d^C(p) = \frac{\alpha}{2(p+1)}$$
$$d^F(p) = \frac{2\alpha}{3(p+1)}$$

As Lemma 3 suggests, $d^C(p) < d^F(p)$ at every p. Under both regimes, equilibrium existence and uniqueness follows from identical arguments, and very similar to those under

[23]The level $d = 0$ can be treated as the minimal level at which the service is usable, but in that case it should also hold that $\ell(0) > 0$.

[24]Committing to their future data-requirements in period 1 would potentially allow Commitment firms to signal their type. A firm that knows it will have higher reputation in the second period finds it more profitable to commit to a higher level of future data sharing, because higher data requirement makes profit more sensitive to current reputation.

exogenous data, because firm profit in period 2 remains increasing in its posterior reputation. An *equilibrium* under the consumer regime, identified using the superscript C, is defined as the unique combination $\{e_k^C, \mu_k^C, d_k^C, \theta_k^C\}$, for $k \in \{1, n, b\}$ such that:

1. Posterior beliefs are consistent with Bayes' rule, given investment level e_1^C.

2. Given e_1^C, active consumers choose data-sharing in each period and state according to $d_k^C = d^C(p_k)$, and $\theta_k^C = u(d_k^C, p_k)$ where $k \in \{1, n, b\}$.

3. $e_n^C = e_b^C = 0$ and e_1^C satisfies the investment f.o.c., given that profits in each $k \in \{1, n, b\}$ are $\Pi^C(p_k)$. Following the derivation of the previous section, the first-order condition is:

$$(1 - \zeta)(\Pi_n^C - \Pi_b^C) = C'(e) \tag{8}$$

where Π_b^C is used as shorthand notation for $\Pi^C(p_b)$. Similarly, we can define the unique equilibrium under firm-control. It is still the case that the firm cannot influence consumer conjectures about investment in period one. Hence, optimal e_1, as well as the endogenous choices of d_n, d_b in either regime, are independent of the choice of d_1, and we will omit d_1 from most of the discussion below. It is worth noting, that in the equilibrium of either regime, consumers will be *under-sharing* with Commitment types relative to the case of complete information, whenever posterior beliefs are interior. They will also be *over-sharing* with Normal types: not only do consumers entertain the possibility of facing a Commitment type, but also recognize that incomplete information gives Normal types investment incentives, which further raises optimal data sharing.

4.2 Limits on data collection in the two regimes

In this subsection, I will be considering a regulator that can, in period 1, *ex-ante* impose levels of data to be shared by active users in each state of period 2. This means that the regulator can condition data collection on the outcome[25] of period 1. The regulator has no informational advantage over consumers and can only use publicly available information about the firm. For any pair of fixed values (d_n, d_b) and the unique equilibrium they induce, I analyse the effects of marginal changes in d_n or d_b on total consumer surplus. The goal is to perform comparative statics of total CS at the unique equilibrium of each endogenous-data regime, derived in the previous subsection.

[25] Or posterior beliefs about the firm's type.

Starting from any pair (d_n, d_b) and the well-defined equilibrium investment and beliefs that these induce, we totally differentiate CS_2, i.e. expected consumer surplus in period 2, to get:

$$\frac{dCS_2}{dd_b} = \frac{\partial CS_2}{\partial e_1}\frac{\partial e_1}{\partial d_b} + \frac{\partial CS_2}{\partial d_b} \tag{9}$$

and the same expression holds for d_n. Each data variable has two effects on CS_2, a *direct*, via changing utility consumers derive from data in period 2, and an *indirect* via changing first-period investment of the Normal type, which depends on the slope of equilibrium investment. I will analyse each of the three terms of (9) separately.

We have defined $p_b = p(\mu_b, 0) = \zeta + (1-\zeta)(1-\mu_b)$ as consumers' expected probability of a breach occurring in the second period, after it has already occurred in the first. This leads to a unique $d^C(p_b)$, endogenously determined in equilibrium. By the assumed quasi-concavity of consumers' utility in d, the sign of the direct effect depends on the comparison of the specific d_b with the value $d^C(p_b)$. Similarly, an increase in d_n benefits consumers via the direct effect iff $d_n < d^C(p_n)$. We summarize the discussion into a Lemma.

Lemma 4 *At the equilibrium induced by a pair of parameters (d_n, d_b), the direct effect on CS_2 of an increase in d_b is negative if and only if $d_b > d^C(p_b)$, while the direct effect of an increase in d_n is negative if and only if $d_n > d^C(p_n)$. If $d_n = d_n^F$ and $d_b = d_b^F$, both direct effects are negative.*

The second part of the Lemma simply echoes Lemma (3). Next, I discuss the investment slope terms. Following the same arguments as in the comparative statics exercises of Lemma 1, we need to consider the sign of the mixed partial derivative of total expected profit with respect to e and d_b (or d_n), holding consumer conjectures fixed. For d_b, this is given by:

$$\frac{\partial^2 \mathrm{T}\Pi}{\partial e_1 \partial d_b} = -(1-\zeta)\frac{\partial \Pi(d_b, p_b)}{\partial d_b} \tag{10}$$

Holding consumer beliefs fixed, a change in d_b will only affect firm incentives via the post-breach profit Π_b. In turn, (10) has the **opposite** sign of $\frac{\partial \Pi(d_b, p_b)}{\partial d_b}$, which by quasi-concavity of the profit function in d is positive if and only if $d_b < d^F(p_b)$, mirroring the argument used above Lemma 4 This is what the Figure shows, for the case of $\zeta = 0$: as d_b approaches the firm optimum from below, the profit from achieving low reputation increases, which implies that incentives to avoid a low reputation decrease. The same argument holds for d_n, except that increases in Π_n *increase* investment incentives. Whether

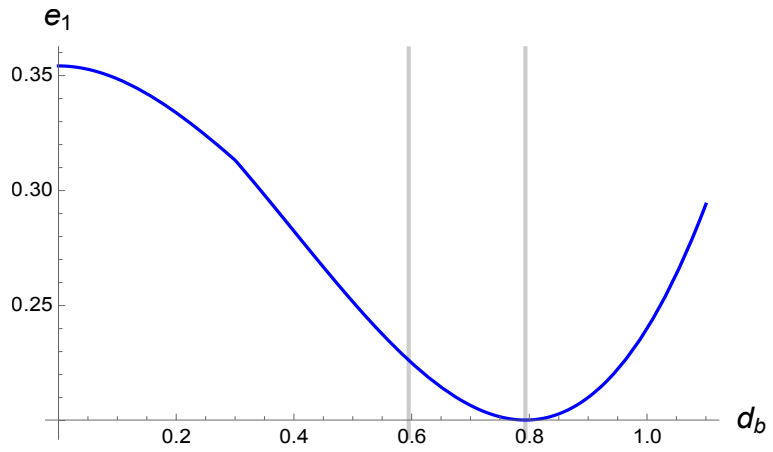

Figure 1: Illustration of Lemma 5, for $\zeta = 0$. The vertical line that crosses the flat part of the curve corresponds to $d^F(1)$, while the one on the left corresponds to $d^C(1) < d^F(1)$, at which point investment is always decreasing in d_b. Drawn for: $c = 3, \alpha = 2.38, d_n = 1.1, \mu = 0.38$.

a marginal change in the exogenous d_n increases equilibrium investment, depends only on whether d_n is larger or smaller than the endogenous $d^F(p_n)$ at the original equilibrium. We state the following Lemma, which is proven in the Appendix.

Lemma 5 *At an interior equilibrium induced by a pair (d_n, d_b), the partial derivative of equilibrium investment with respect to d_b, $\partial e^*/\partial d_b$, is positive if and only if $d_b > d^F(p_b)$, with equality at $d_b = d^F(p_b)$. Similarly, $\partial e^*/\partial d_n$ is positive if and only if $d_b < d^F(p_b)$. At the equilibrium values of the firm regime, there are no first-order effects on investment.*

It is helpful to note that, for $\zeta = 0$, ~~the~~ perfect-bad news ~~assumption~~ simplifies this argument because p_b is always equal to 1, independent of other parameters, hence $d^C(p_b)$ becomes a scalar[26], $d^C(1)$. For this particular case, Figure 1 illustrates Lemma 5.

Now that we have understood the direction of effects on equilibrium investment, we ask how these will translate into changes in expected consumer surplus. Starting with second period consumer surplus, we turn to the effect of higher investment e_1 by the Normal type on *second-period* consumer surplus $\frac{\partial CS_2}{\partial e_1}$. Differentiating CS_2 with respect to e_1 and evaluating at the *consumer-regime equilibrium*, we obtain:

$$\frac{\partial CS_2}{\partial e_1} = (1-\zeta)(1-\mu)\left[\int_0^{\theta_n^C}\Big(u(d_n^C,1)-\theta\Big)d\theta - \int_0^{\theta_b^C}\Big(u(d_b^C,1)-\theta\Big)d\theta\right]$$

[26]Even for $\zeta = 0$, d_n^C would remain endogenous.

e_1 affects CS_2 in two ways: directly, via changing the frequency with which breaches occur and thus the distribution of firm posterior reputations, holding the latter fixed. Indirectly, by affecting posterior reputations μ_n, μ_b and thus participation decisions. However, there is no first-order impact of those on CS_2 because there are no externalities in this model; consumers choose participation optimally, given their information and beliefs, which in equilibrium are correct. The only first-order effect is the direct, holding μ_n, μ_b fixed.

I will refer to the above as the *signal-jamming* effect of investment and I claim that at the consumer-regime equilibrium, it is negative. Each integral in the above expression is over consumer net utilities conditional on facing a Normal type, i.e. with second-period probability of breach equal to 1. This conditional consumer surplus is maximized by the combination $d^C(1)$ and the quasi-concavity of utility means that $u(d^C(p_n), 1) < u(d^C(p_b), 1) < u(d_b^C(1), 1)$, so that the integral on the right is always (weakly) larger[27].

Intuitively, conditional on facing a Normal type, a breach reveals the firm's type more accurately and helps consumers make better decisions in period 2. As long as lack of a data breach remains *informative* and consumers *act* on that information, i.e. as long as consumers participation and data-sharing decisions are different in the two states of period 2, higher investment by the Normal type means that consumers are more frequently misguided into giving away more data than they would, if they knew for a fact that they are facing a Normal type. Notice that for $e_1 \to 1$, first-period outcomes are no longer informative on the firm's type and the signal-jamming effect becomes zero.

Lemma 6 *The signal-jamming effect of investment on second-period consumers surplus is always negative when evaluated at the equilibrium of the game with consumer control.*

Finally, given the last few results, we can sign the total derivatives of CS_2 at each of the regime's unique equilibrium. Consider changes in either parameter, starting from the equilibrium values of the equilibrium under ex-post control by the firm. By Lemma 5, investment is unchanged by local changes to either data term, so only the direct loss to consumers remains.

Corollary 1 *From the equilibrium under firm control, a planner can raise total consumer surplus by imposing a marginal reduction in the amount of data that firms with either high or low reputation can ask for in period two.*

[27]Their problem is exacerbated by the fact that more consumers regret their participation ex-post following a "no breach" than a "breach" realization, i.e. $\theta_n > \theta_b > \theta(d^C(1), 1)$.

In other words, the planner can raise total consumer surplus by imposing small caps on the amount of data that firms can collect in period 2, relative to the equilibrium of the firm regime. This is *total* consumer surplus, since d_n, d_b only affect CS_1 via investment and neither has a first-order impact on e_1 at the examined equilibrium. Consumers benefit from less data-sharing in period 2 because, by Lemma 4, firms ask too much data in the firm-control equilibrium.

On the other hand, looking at the total derivatives at the consumer-control equilibrium:

$$\left.\frac{dCS_2}{dd_b}\right|_{(d_n^C, d_b^C)} = \underbrace{\frac{\partial CS_2}{\partial e_1}}_{(-)}\underbrace{\frac{\partial e_1}{\partial d_b}}_{(-)} + \underbrace{\frac{\partial CS_2}{\partial d_b}}_{=0} > 0 \tag{11}$$

$$\left.\frac{dCS_2}{dd_n}\right|_{(d_n^C, d_b^C)} = \underbrace{\frac{\partial CS_2}{\partial e_1}}_{(-)}\underbrace{\frac{\partial e_1}{\partial d_n}}_{(+)} + \underbrace{\frac{\partial CS_2}{\partial d_n}}_{=0} < 0 \tag{12}$$

The difference driving the results is that an increase in d_n from d_n^C will increase investment (revenue with high reputation increases), whereas an increase in d_b will decrease investment (revenue with low reputation increases). Thus, one change causes more and the other causes less *signal jamming*, and neither has first-order *direct* effects. We learn that locally, CS_2 can increase by committing to a larger d_b, i.e. consumers punish too hard and share too little data with low-reputation firms, but smaller d_n, they give out too much data to high-reputation firms[28].

Corollary 2 *From the equilibrium under ex-post control of the consumers, the planner can increase CS_2 by imposing small caps on data sharing for high-reputation firms, i.e. by decreasing d_n, but not for low-reputation ones.*

These corollaries are meant to relate the model in this paper with GDPR-style regulation regarding opt-out rights of consumers, which is interpreted here as consumers choosing how much of their data to share with the firm[29], given what they believe about the cyber-security of the firm. Corollary 1 tells us that some degree of opt-out rights is always beneficial to consumers, relative to the setting where firms impose make take-it-or-leave-it offers to consumers with respect to data collection. Corollary 2 suggests that CS_2 can increase relative to the "opt-out" equilibrium, by imposing *additional restrictions* on data collection by firms that are perceived as "low-risk", i.e. those with a good track record

[28] I have been silent about the impact of d_1 on welfare. In the $T = 2$ model, investment incentives are not affected by d_1, so that the planner cannot do better than d_1^C.

[29] Depending on the interpretation of d; see also the discussion at the end of Section 3.

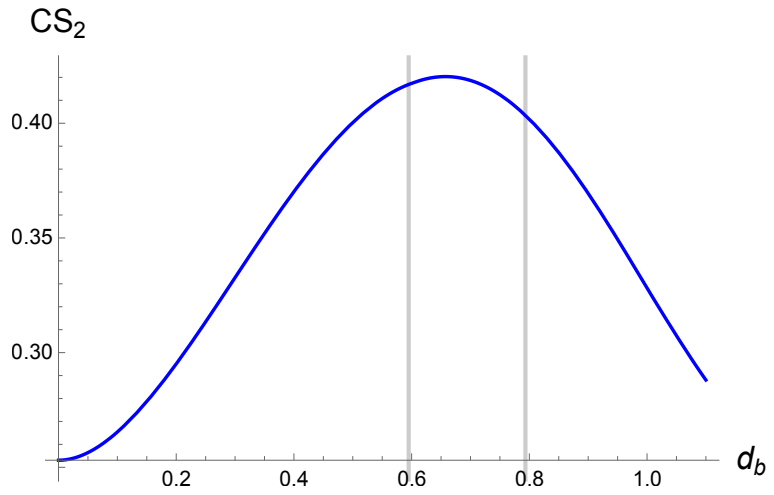

Figure 2: Fixing some d_n, we plot CS_2 as a function of d_b. Between the left and right vertical lines, which correspond to $d^C(1)$ and $d^F(1)$, respectively, there exists a local maximum of CS_2: for the given functional form assumptions, it is a global maximum. Drawn for: $c = 11, \alpha = 2.38, d_n = 1.1, \mu = 0.38$ and $\zeta = 0$.

in cyber security. On the other hand, regarding data collection by "high-risk" firms, the planner should restrict consumers' ability to opt-out, for instance by specifying some information that active consumers have to share.

In Figure 2, I focus on the simpler case of $\zeta = 0$, in which, as I have discussed, both $d^C(1), d^F(1)$ are specific scalars[30]. The analysis implies that for any fixed d_n, there exists a local maximum of second-period consumer surplus at some $d_b^{**} \in (d^C(1), d^F(1))$. Of course, so far I have ignored that changes around the consumer-control equilibrium will also induce changes to consumer surplus of period 1. The latter, is defined as:

$$CS_1 = \int_0^{\theta_1} \Big(u(d_1, p_1) - \theta \Big) d\theta \tag{13}$$

so that the total derivative with respect to e_1 is:

$$\frac{dCS_1}{de_1}\Big|_{(d_n^C, d_b^C, d_1^C)} = \theta_1^C \frac{\partial u(d_1, P_1^b)}{\partial e_1} > 0 \tag{14}$$

Once again, effects via changing θ_1 are not of first-order; consumers make activity decisions optimally in equilibrium. In addition, the first-order effects from changing d_1 are zero at d_1^C. The purely positive impact of investment on CS_1 then comes from a lower

[30]The case with $\zeta > 0$ is more complicated because as we vary d_b, holding d_n fixed at some value, p_b changes at every equilibrium. Hence, $d^C(p_b), d^F(p_b)$, too and we must track how these compare to the current value of d_b in order to sign the direct and indirect effects, as suggested by Lemmas 4 and 5.

breach probability[31]. So how can we compare the induced increase to CS_1 due to fewer breaches with the decrease in CS_2 induced by less learning? Observe that at high levels of investment, θ_1 is larger, **and** d_1^C should be larger too, because consumers expect their data to be safer.

$$\frac{\partial u(d_1, P_1^b)}{\partial e_1} = (1-\mu)(1-\zeta)\ell(d_1^C) \tag{15}$$

The above derivative is increasing d_1^C, and thus larger at higher levels of investment. This is intuitive; when consumers expect fewer breaches and share a lot of data with the firm, the marginal utility of further reduction in the probability of a breach increases. On the other hand, the signal-jamming effect, via which CS_2 changes, shrinks towards zero[32] at high levels of investment. We take the second order partial derivative to see that:

$$\frac{\partial^2 CS_2}{\partial e_1^2} = (1-\zeta)(1-\mu)\left[\underbrace{\frac{\partial \theta_n}{\partial e_1}}_{(-)}\underbrace{\left(u(d_n^C,1)-\theta_n^C\right)}_{(-)} - \underbrace{\frac{\partial \theta_b}{\partial e_1}}_{(+)}\underbrace{\left(u(d_b^C,1)-\theta_b^C\right)}_{(-)}\right] > 0 \tag{16}$$

Participation is lower because e_1 lowers the high posterior reputation. Conditional on facing a Normal type, equilibrium participation levels following no-breach the equilibrium marginal consumer always has ex-post regret. Intuitively, as $e_1 \to 1$, the signal-jamming effect should shrink, because as $e_1 \to 1$, it implies $\mu_n \to \mu$, bringing θ_n^C and d_n^C closer to their state b corresponding quantities. Essentially at higher e_1, consumers understand that a lack of data-breach is often caused by a Normal type and are cautious with participation and data-sharing in the second period. This means that ex-post regret is of lower magnitude in period 2 when a Normal type achieves high reputation. This is illustrated in Figure 3: I plot how the partial derivative of CS_2 with respect to e_1 (i.e. the signal-jamming effect) changes with e_1, under ex-post consumer-optimal decisions. As e_1 varies, consumer posterior beliefs change in each state and so do the equilibrium levels of data sharing, d_1^C, d_b^C, d_n^C at which I evaluate the partial derivative.

Lemma 7 *When parameters are such that the consumer-control equilibrium features higher levels of* e_1^C*, the negative signal-jamming effect is of lower magnitude,* $\frac{\partial^2 CS_2}{\partial e_1^2} > 0$*. At the same time, the positive impact of investment on* CS_1 *is even higher,* $\frac{\partial^2 CS_1}{\partial e_1^2} > 0$*. As a*

[31] When there is also an effect via d_1, and given that $\theta_1 = u(d_1, P_1^b)$, we observe that CS_1 increases iff the participation cutoff does. This is the case because $u_{\theta\delta} = 0$: whenever *any* consumer benefits by the joint change in d_1 and e_1, the marginal consumer benefits too, thus the location of the margin shifts up. For our baseline functional form, greater investment increases equilibrium CS_1.

[32] If $\zeta > 0$; otherwise, the first-period outcomes are always informative and the two posteriors always remain bounded away from each other.

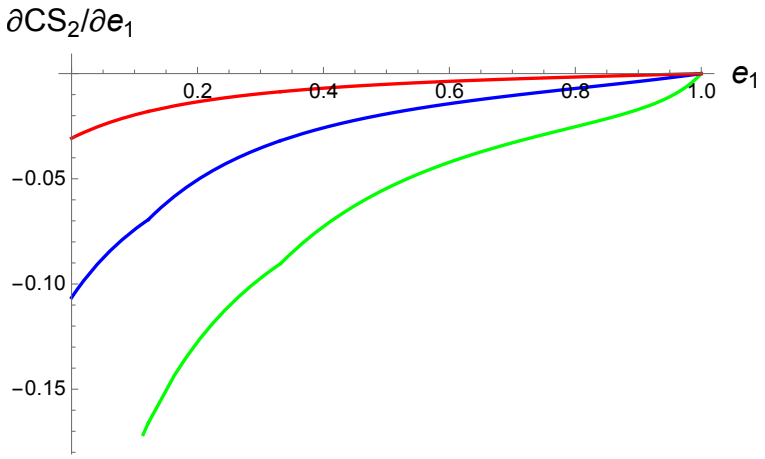

Figure 3: Vertical axis is $\partial CS_2/\partial e_1$ evaluated at $d_n = d^C(\mu_n(e_1)), d_b = d^C(\mu_n(e_1))$. Higher equilibrium investment means lower magnitude of signal-jamming. Drawn for: $\alpha = 2.38, \mu = 0.38$, and $\zeta \in \{0.1, 0.3, 0.5\}$. Lower ζ corresponds to a curve higher up. For $\zeta > 0$, all curves converge to the horizontal axis.

result, it is more likely that starting from high e_1^C*, increases in* e_1 *are more likely to raise total consumer surplus.*

5 The model in steady state

In this section, I extend the model to an infinite horizon variation. Specifically, I will focus on the regime of *consumer control* over data sharing. I will make appropriate assumptions, so that there is a stationary equilibrium in which the Normal type uses the same e^* in every period. The objective it two-fold: first, to construct a model in which all insights derived in the previous section extend to; the qualitative insight of the model does not depend on the lack of reputational concerns in the last period of a finite-horizon model. Second, to use this model in order to integrate the security and learning effects of cyber-security and identify a consumer-surplus maximizing policy. In this model of $T = \infty$, the firm lives forever, and has private knowledge of its time-invariant type. The timing is as follows:

1. A new consumer cohort is born in every period, t. Once they become alive in period t, they immediately learn the security *outcome* of period $t-1$, and only that.

2. Based on that, they update to their beliefs are either μ_n or μ_b.

3. Consumers choose whether to participate[33] and how much data to share with the firm.

4. A Normal type firm chooses the unobserved e_t in every period.

5. A breach occurs or not, affecting the current cohort of active consumers. After that, they exit the game.

Each cohort of consumers holds the same prior μ over the firm's type and the same conjecture $\tilde{e}$ over the investment made *in every period* by the Normal type. This conjecture affects both their belief about the firm's type based on the outcome of last period, as well as their perceived probability with which a Normal type will suffer a breach in the current period. Same as in the $T = 2$ model, investment incentives for the firm in period t come from the benefit of achieving a high reputation in period $t+1$. Thanks to the assumptions we have made above and in the presentation of the benchmark model of Section 3, the firm only considers the expected revenue of $t + 1$ when choosing e_t. The first-order condition for investment looks identical to (8), and equilibrium must satisfy the conditions already discussed. The main difference to (8) is in specifying the probability with which consumers of cohort t expect the firm to suffer a breach following each outcome of $t - 1$. Following a breach:

$$\begin{aligned} p^{\infty}(\mu_b, e) &:= \mu_b(e,\zeta)\zeta + (1-\mu_b(e,\zeta))(\zeta + (1-\zeta)(1-e)) \\ &= \zeta + (1-\zeta)(1-\mu_b(e,\zeta))(1-e) \end{aligned} \tag{17}$$

and having observed "no-breach" in $t - 1$, consumers expect to be breached in t with probability:

$$p^{\infty}(\mu_n, e) = \zeta + (1-\zeta)(1-\mu_n(e))(1-e) \tag{18}$$

Again, μ_n does not depend on ζ because the latter does not affect the LR conditional on "no-breach". In addition:

$$\frac{\partial p^{\infty}(\mu_b, e)}{\partial e} = -(1-\zeta)\left[(1-\mu_b) + (1-e)\frac{\partial \mu_b}{\partial e}\right] < 0 \tag{19}$$

[33]Participation decisions add no qualitative insight at this point and I could also assume that all consumers have an outside option equal to zero. I maintain the assumption of incomplete participation for continuity with the previous sections.

while the change in $p^\infty(\mu_n, e)$ has *ambiguous* sign because the high posterior decreases in e.

$$\frac{\partial p^\infty(\mu_b, e)}{\partial e} = -(1-\zeta)\left[(1-\mu_n) + (1-e)\underbrace{\frac{\partial \mu_n}{\partial e}}_{(-)}\right] \tag{20}$$

This fact can potentially cause issues in equilibrium uniqueness; the argument relied upon in the $T = 2$ model rested on the marginal benefit of investment *decreasing* in consumers' beliefs about investment. A sufficient condition for this is that $p^\infty(\mu_n, e)$ is *increasing* in e. As $e \to 0$, $\mu_n \to 1$ and thus $p^\infty(\mu_n, e)$ is unambiguously **increasing** in e, so that the standard intuition prevails and equilibrium marginal benefit is decreasing. Since $p^\infty(\mu_n, e)$ is increasing in an interval of e including $e = 0$, marginal benefit is decreasing in e at $e = 0$. This ensures existence of a solution for the first-order condition[34]

On the other hand, as $e \to 1$, the derivative becomes unambiguously *negative*. With very high effort of the Normal type, the change in beliefs about the type, $\frac{\partial \mu_n}{\partial e}$, becomes less important in determining the total change in the prob. of a breach[35]. It is worth noting that the posterior $p^\infty(\mu_n, e)$ is **less** sensitive to e than $p^\infty(\mu_b, e)$, simply because the prob of a Normal type is lower. Thus, even when the former is decreasing in e, it does so at a rate slower than $p^\infty(\mu_b, e)$, which is a force driving the decreasing marginal benefit of investment and uniqueness of equilibrium we observe numerically.

Proposition 2 *A stationary Markov Perfect Equilibrium of the steady-state game exists, in which the Normal type firm plays $e^* < 1$ in every period. It is unique for sufficiently convex cost of investment.*

Next, I ask what is the time-invariant combination (d_n, d_b, e) that maximizes CS^∞, i.e. discounted expected consumer surplus in this game. Notice that when we fix the Normal type's investment at e, the data sharing rule that maximizes CS^∞ must be given by the ex-post optimal rule: In a model without consumption or data externalities, the only value to deviating from the ex-post optimal decisions of each consumer would be to influence the firm's strategic incentives. I am thus looking for the optimal value of e to induce, for instance via employing an (expected) penalty f that a firm incurs each time it suffers a

[34] And at a point that is indeed a local maximum for the firm's problem.

[35] We obtain the **negative** sign whenever:

$$\frac{1-\mu}{\mu}e^2 + 2e - 1 > 0 \tag{21}$$

which happens whenever $e > \frac{\sqrt{\mu}}{1+\sqrt{\mu}}$. This lower bound is smaller than 0.5 and it is *independent* of ζ.

data breach[36]. It is easy to show that[37]:

Lemma 8 *Assume a penalty $f > 0$ is imposed to a firm every time it suffers a breach. If, for every f, a unique stationary equilibrium investment exists, it is increasing in f.*

Ignoring participation constraints of the firm, due to potentially negative, i.e. assuming large enough revenue per consumer or low enough investment cost, the planner can thus implement the optimal value of CS^{∞} in equilibrium by giving consumers ex-post rights in choosing data collection levels, and using a penalty to fine-tune the firm's investment incentives.

Proposition 3 *The tuple $(d_n = d^C(\zeta), d_b = d^C(\zeta), e = 1)$ maximizes discounted expected consumer surplus, CS^{∞}.*

When N fully imitates C, consumers do not value learning. Hence, maximum security and the associated ex-post optimal data-sharing choices jointly maximize consumer surplus. When both types us $e = 1$ and breaches occur with probability ζ, the two types become indistinguishable, hence $d^C(p_n) = d^C(p_b) = d^C(\zeta)$ and the expected CS induced by this tuple following either b or n is $CS^{\infty}(\zeta, d^C(\zeta))$. make sure you define this function above Of course, this insight would also apply to the $T = 2$ model.

But what if the planner *cannot* impose a large enough fine such that $e = 1$ in equilibrium? Will increasing the penalty, and thus equilibrium investment monotonically increase CS^{∞}? In order to answer this, I examine the shape of CS^{∞} with respect to e, accounting for the fact that participation and data-sharing choices will be adapting to the changing breach probabilities in each state. By Proposition 3, this attains its global maximum at $e = 1$, at which point it must also be increasing. Assuming for a moment that participation is always complete, the inequality that must be satisfied for negative slope at $e = 0$ is:

$$\underbrace{\ell(d^C(p(\mu_b, 0))}_{\text{security gain}} < \underbrace{\left[v(d^C((p(\mu_b, 0))) - \ell(d^C(p(\mu_b, 0))\right] - \left[v(d^C(0)) - \ell(d^C(0))\right]}_{\text{Gain from more accurate signal about firm's type}} \tag{22}$$

[36] I am thinking about a strict liability regime in which the regulator cannot ex-post verify whether the breach occurred due to the firm's negligence ($e < 1$) or due to the exogenous shock ζ which the firm cannot protect against. Under a large value of ζ, a large penalty would potentially make a firm's profit drop to zero, but I am ignoring potential violations of participation constraints at the moment.

[37] In an earlier section, it is mentioned that the model can accommodate $\hat{e} < 1$ to be played by the Commitment type in each period. A more detailed discussion of the above Lemma when this is the case is relegated to the Appendix.

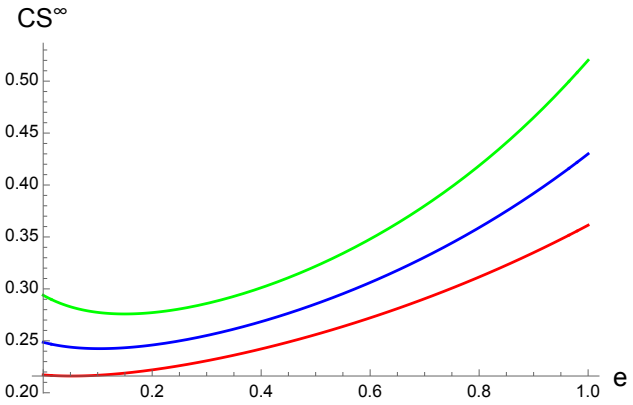

Figure 4: As we exogenously vary the level of investment of the Normal type, consumers adapt their beliefs and optimal participation and data decisions. The green, blue and red curves correspond to $\zeta \in \{0, 0.1, 0.2\}$ respectively. Drawn for $\alpha = 2.2, \mu = 0.42$. For sufficiently small ζ, the curve is downward sloping at $e = 0$. Intuitively, learning about the type of the firm is more important for smaller ζ.

At $e = 0$, the Normal type is always breached, in which case consumers share $d^C(p(\mu_b, 0))$and that a good outcome is perfectly revealing of a Commitment type. The difference in the right-hand side is between consumer utility when $d^C(p(\mu_b, 0))$ is shared and when $d^C(0)$ is shared, given that the probability of a breach is 1, which is the case when $e = 0$ and the firm is Normal type. This difference must be positive by quasi-concavity of u. At $e = 0$, the difference between posteriors is maximized, and so is the right-hand side. At the same time, the belief $p(\mu_b, e)$ is maximized, hence the left-hand side is minimized. This is the intuition of Lemma 7. We show in the Appendix that if (22) is satisfied for the case of complete participation in every state, it is also satisfied for the case of consumer-optimal participation. The intuition remains the same.

6 Conclusion

In a two-period model, I examine the incentives of a digital service monopolist to invest in unobserved data security, when it charges no access fees but instead monetizes consumer data. Consumers suffer privacy-related disutility when data-breaches occur, and the firm wants to earn a reputation for protecting users' data to maintain high activity in period two. I introduce two regimes of endogenous data sharing: in the regime of firm control, data-sharing requirements are chosen by the firm in every period to maximize current profits. If it is consumers, data-sharing is chosen to maximize current-period consumer

surplus, accounting for the firm's reputation. I ask whether a social planner can improve ex-ante consumer surplus by committing to different levels of data sharing in period two, relative to the regulation-free equilibria, and I allow data sharing to depend on the firm's posterior reputation.

Ex-ante commitment to data sharing affects consumer surplus directly, but also via equilibrium investment. Starting at the firm-control equilibrium, the effects on investment are dominated, and the planner can improve total CS by reducing the amount of data that both high and low reputation firms collect. On the other hand, compared to the ex-post consumer optimum, committing to less data sharing following a breach induces higher security; the ex-ante optimal level trades-off higher security and more "signal jamming": greater investment impedes learning about the true levels of cyber risk which harms consumers in the second period.

References

Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The economics of privacy. *Journal of Economic Literature*, 54(2):442–92, 2016. URL https://EconPapers.repec.org/RePEc:aea:jeclit:v:54:y:2016:i:2:p:442-92.

Toni Ahnert, David Cimon, and Ryan Riordan. Cyber security and ransomware in financial markets. *CEPR Press Discussion Paper No. 17403.*, 2022. doi: https://cepr.org/publications/dp17403.

Guy Aridor, Yeon-Koo Che, and Tobias Salz. The effect of privacy regulation on the data industry: empirical evidence from GDPR. *RAND Journal of Economics*, 54(4):695–730, December 2023. doi: 10.1111/1756-2171.12455. URL https://ideas.repec.org/a/bla/randje/v54y2023i4p695-730.html.

Alexandre de Cornière and Greg Taylor. A Model of Information Security and Competition. TSE Working Papers 21-1285, Toulouse School of Economics (TSE), December 2021. URL https://ideas.repec.org/p/tse/wpaper/126391.html.

Anastasios Dosis and Wilfried Sand-Zantman. The ownership of data. *The Journal of Law, Economics, and Organization*, 39(3):615–641, November 2023. doi: 10.1093/jleo/ewac001. URL https://doi.org/10.1093/jleo/ewac001.

Itay P. Fainmesser, Andrea Galeotti, and Ruslan Momot. Digital Privacy. *Management Science*, 69(6):3157–3173, June 2023. doi: 10.1287/mnsc.2022.4513. URL https://ideas.repec.org/a/inm/ormnsc/v69y2023i6p3157-3173.html.

Avi Goldfarb and Catherine Tucker. Introduction to The Economics of Privacy . May 2023. URL https://ideas.repec.org/h/nbr/nberch/14786.html.

Bengt Holmström. Managerial incentive problems: A dynamic perspective. *Review of Economic Studies*, 66(1):169–182, 1999.

Bruno Jullien, Yassine Lefouili, and Michael Riordan. Privacy Protection, Security, and Consumer Retention. *TSE Working Papers*, August 2020. URL https://ideas.repec.org/p/tse/wpaper/32902.html.

Shinichi Kamiya, Jun-Koo Kang, Jungmin Kim, Andreas Milidonis, and René M. Stulz. Risk management, firm reputation, and the impact of successful cyberattacks on tar-

get firms. *Journal of Financial Economics*, 139(3):719–749, 2021. URL https://EconPapers.repec.org/RePEc:eee:jfinec:v:139:y:2021:i:3:p:719-749.

Pantelis Koutroumpis, Farshad Ravasan, and Taheya Tarannum. (under) investment in cyber skills and data protection enforcement: Evidence from activity logs of the uk information commissioner's office. *Working Paper*, July 2022. URL https://ssrn.com/abstract=4179601.

Yassine Lefouili, Leonardo Madio, and Ying Lei Toh. Privacy regulation and quality-enhancing innovation. *The Journal of Industrial Economics*, 2024. doi: https://doi.org/10.1111/joie.12374. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/joie.12374.

Tesary Lin. Valuing intrinsic and instrumental preferences for privacy. *Marketing Science*, 41, 2022. URL https://doi.org/10.1287/mksc.2022.1368.

Sarit Markovich and Yaron Yehezkel. "for the public benefit": who should control our data? Working Papers 21-08, NET Institute, 2021. URL https://EconPapers.repec.org/RePEc:net:wpaper:2108.

A Benchmark model

A.1 Proposition 1

Similar to the statement of the Proposition, we take cases according to the value of ζ. If $\zeta = 1$, then there is no value to investment and $e^* = 0$. Consumers know they will suffer a breach in every period and make participation decisions accordingly. If $1 > \zeta > 0$, bad news are imperfect. In that case, $e = 1$ cannot be supported in equilibrium, because $\mu_n(1) = \mu_b(1) = \mu$, hence consumer decisions in period 2 are the same following either outcome of period 1 and the marginal benefit of investment becomes zero. The first-order condition that defines an equilibrium with interior investment, $e_1 < 1$, is given by:

$$(1-\zeta)\Big(\Pi(d, p(\mu_n(e_1), 0)) - \Pi(d, p(\mu_b(e_1), 0))\Big) = C'(e_1) \tag{23}$$

and notice that, in equilibrium, e_1 enters in the firm's marginal cost and in consumers' posterior beliefs. Every solution to this equation constitutes an equilibrium; consumers hold correct beliefs and the firm's decision satisfies the relevant first-order condition. Since

the firm's maximization problem takes beliefs as given, the firm's marginal benefit of investment is flat and the second-order condition is always satisfied for convex cost.

It it easy to see that the equilibrium marginal benefit curve (the right-hand side of (23)) is *decreasing* in e_1. Since $\Pi = r(d)F(u(d,p))$ and $u(d,p) = v(d) - p\ell(d)$, we know that $d\Pi/dp < 0$ and differentiating the l.h.s., which I define as the function G of investment and parameters, yields:

$$\frac{\partial G}{\partial e_1} = (1-\zeta)\Big[\overbrace{\frac{\partial \Pi(d, p(\mu_n(e_1),0))}{\partial p}}^{(-)} \overbrace{\frac{\partial p(\mu_n(e),0)}{\partial e}}^{(+)} - \underbrace{\frac{\partial \Pi(d, p(\mu_b(e_1),0))}{\partial p}}_{(-)} \underbrace{\frac{\partial p(\mu_b(e),0)}{\partial e}}_{(-)}\Big] < 0$$

which is clearly *negative*, because μ_b increases in e_1, and the opposite for μ_n. Since for all $e_1 < 1$, $\mu_n(e_1) > \mu_b(e_1)$, the left-hand side is positive as long as $\Pi(d, p(\mu_n(e_1),0)) > 0$, i.e. as long as a positive mass of consumers uses the service when beliefs are $p(\mu_n(e_1),0)$. Under this assumption, at $e_1 = 0$, the left-hand side is positive, larger than $C'(0) = 0$. We have already argued that the left-hand side tends to zero as $e_1 \to 1$, in which case it is smaller than the right-hand side. By continuity of both sides in e_1 and the Intermediate Value Theorem, a solution must exist, strictly in the interior of $(0,1)$.

The case of $\zeta = 0$, i.e. perfect bad news proceeds similarly. Even for constant $\mu_b = 0$, the equilibrium marginal benefit is a decreasing function of e_1 and there are two possibilities. Either $\Pi(d, p(\mu_n(e_1),0)) \leq 0$ and $e^* = 0$, which I assume away as discussed above, or $\Pi(d, p(\mu_n(e_1),0)) > 0$: in this case, there is an interior equilibrium iff $\Pi(d, p(\mu_n(1),0)) - \Pi(d, p(0,0)) < C'(1)$, which is guaranteed by $\lim_{e_1 \to 1} C'(e_1) = \infty$. If the marginal cost does not increase fast enough, $e^* = 1$.

A.2 Lemma 1

To show the first result, I appeal to the Implicit Function Theorem, to implicitly differentiate equation (23) that defines an interior equilibrium. At parameter values that yield an equilibrium with $e^* < 1$, equilibrium investment is a differentiable function of d and it holds that:

$$\begin{aligned}\frac{\partial e}{\partial d} &= -\frac{G_d}{G_e} = \Big(\frac{(1-\zeta)}{-G_e}\Big)\frac{\partial}{\partial d}\Big[\Pi(d, p(\mu_n(e_1),0)) - \Pi(d, p(\mu_b(e_1),0))\Big] \\ &= \Big(\frac{1}{-G_e}\Big)\int_{p(\mu_b(e_1),0)}^{p(\mu_n(e_1),0)} \frac{\partial^2 \Pi(d,p)}{\partial d \partial p} dp > 0\end{aligned}$$

because $p(\mu_n(e_1), 0) < p(\mu_b(e_1), 0)$ and as argued in the proof of Proposition 1, $G_e < 0$ and $\frac{\partial^2 \Pi(p,d)}{\partial d \partial p} = r'(d)\frac{\partial u(p,d)}{\partial p} + r(d)\frac{\partial^2 u(p,d)}{\partial d \partial p} < 0$. For this derivation, I used the uniform F assumption.

B Endogenous data collection

B.1 Lemma 3.

For the first part of the statement I will repeat the verbal argument from the main text: if the firm finds it optimal to ask for $d^F(p) < d^{max}$, it must be that a further increase in d will reduce consumer utility and thus the mass of active users. Otherwise, an increase in d would surely raise revenue, yielding a contradiction. For the second part, assuming that both $u(d, p)$ and $\Pi(d, p)$ are differentiable and quasi-concave in d for all $p \in (0, 1)$, the respective maximizers are uniquely determined as solutions to first-order conditions. The firm's marginal benefit of increasing d is:

$$\frac{\partial \Pi(p, d)}{\partial d} = r'(d)F(u(d, p)) + r(d)f(u(d, p))\frac{\partial u(d, p)}{\partial d}$$

if the consumers choose $d^C(p) = d^{max}$, it must be that their marginal utility is still weakly positive, implying that the above derivative is also positive, so that $d^F(p) = d^{max}$, too. If $d^C(p) < d^{max}$, then marginal utility must be zero by the first-order condition, the above derivative is again positive, and $d^C(p) < d^F(p)$ holds as a strict inequality.

C Model in steady state

C.1 Claim about incomplete participation

Claim: When consumers have no outside options, or $u(d^C(1), 1)$ is large enough, and everyone always uses the service, the slope is negative at $e_1 = 0$, i.e. (22) is satisfied. Then, the slope of CS with respect to e_1 as $e_1 \to 0$ is also negative whenever $\theta_n = \theta_n^C$ and $\theta_b = \theta_b^C$ are smaller than 1.

By the usual envelope argument, when taking the derivative of CS with respect to e_1, there are no first-order effects from changes in either participation or data-sharing. Hence,

the condition for negative slope at $e_1 = 0$ becomes:

$$\theta^C(p(\mu_b,0))\ell(d^C(p(\mu_b,0))) < \int_0^{\theta^C(p(\mu_b,0))} \Big[v(d^C(p(\mu_b,0))) - \ell(d^C(p(\mu_b,0))) - \theta\Big] f(\theta)d\theta - \int_0^{\theta^C(0)} \Big[v(d^C(0)) - \ell(d^C(0)) - \theta\Big] f(\theta)d\theta$$

and arguments in the main text that the right-hand side is positive. Since $\theta^C(0) > \theta^C(p(\mu_b,0))$, this can be rewritten as:

$$\int_0^{\theta^C(p(\mu_b,0))} \Big[\ell(d^C(p(\mu_b,0))) - (v(d^C(p(\mu_b,0))) - \ell(d^C(p(\mu_b,0))) + (v(d^C(0)) - \ell(d^C(0))\Big] f(\theta)d\theta < - \int_{\theta^C(p(\mu_b,0))}^{\theta^C(0)} \Big(v(d^C(0)) - \ell(d^C(0)) - \theta\Big) f(\theta)d\theta$$

the outside options do not affect optimal data-sharing decisions of active consumers, hence the $d_b(0), d_n(0)$ terms are the same in the inequalities that refer to the cases with and without flexible participation. Hence, the integrand in the left-hand side is **negative** whenever (22) is satisfied, and the right-hand side is always positive: when the true probability of a breach is 1, any consumer with outside option in the range $[\theta^C(p(\mu_b,0)), \theta^C(0)]$ of the integral obtains negative expected utility when facing a Normal type.