Strategic Incentives and Regulation in Cyber Security

Emmanouil Spyridon Perdikakis
Jesus College

Supervisor: Dr Margaret Meyer

Thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

Department of Economics
University of Oxford

Trinity Term 2025

Acknowledgements

I am grateful to my supervisor, Margaret Meyer, for her continuous support during my DPhil. I thank her for encouraging me to explore every direction of my research interests and for our lengthy, detailed conversations on each of my projects. I have left every one of our discussions a better economist.

I also thank Alexei Parakhonyak and Greg Taylor for their generous support and feedback, from which my work has greatly benefited. My research collaboration with Greg, in particular, has been one of the most exciting parts of my DPhil.

For their close friendship, everyday support, and countless discussions on economics and research, I thank Carlos Akkar, Lukas Boehnert, Giulio Gottardo, Amelie Mennerich, Hannah Römer, and Rafael Suchy. I am thankful to Clara Albiñana, for bringing joy and encouragement to the final stage of the DPhil.

I gratefully acknowledge funding from the University of Oxford Department of Economics, and the A.G. Leventis Foundation Scholarship Program.

Finally, and most importantly, I am deeply grateful to my family for their unwavering support during the past few years.

Table of Contents

List of Figures

List of Tables

Abstract

This thesis studies the optimal design of policy in the domains of data protection and cyber security. Each of the three main chapters consists of an independent paper. The first paper studies a digital platform's incentives to invest in protecting the consumer data it collects. Data security investment is unobserved by consumers and incentives are reputational. In a two-period model, I show that a planner can raise total consumer welfare by imposing ex-ante limits on data collection based on a firm's history of data breaches. The optimal policy depends on whether firms or consumers control data collection in each period. Additionally, I use the model to evaluate established policies: minimum security standards and limits to data retention. The second paper studies how markets for ransomware insurance affect the welfare of firms and hackers, and asks whether regulation can improve outcomes. In the model, hackers may or may not observe victims' insurance contracts, and firms may be unable to pay ransom due to liquidity constraints. Insurance has commitment value for the firms in their bargaining with hackers and can reduce ransom demands. Regulatory caps on insurance for ransom payments guarantee that the presence of insurers makes firms better off, and hackers worse off. The third paper focuses on a fundamental trade-off that regulators face when designing data-breach notification laws: high penalties following disclosure encourage firms to invest in cybersecurity ex-ante but also to conceal breaches ex-post. In the model, a firm only discloses a breach after it has become pessimistic about the prospect of concealing it. I characterize the optimal policy for a regulator who can commit to penalties following disclosure of a breach. I examine design of the optimal policy when disclosure delay is ex-post verifiable and

the regulator uses delay-dependent penalties to screen firms' private information.

1 | Introduction

In this thesis, I employ methods of Information Economics and Industrial Organization to analyze policy questions in the domains of data protection and cyber security.

My work is motivated by two types of socially harmful cyber attacks. The first are traditional data breaches, whose purpose is the theft and exploitation of consumers' personal data. The second are the increasingly frequent ransomware attacks, in which malicious parties demand ransom payments from victim firms, under the threat of causing severe business interruption and leaking customers' data.

In Chapters 2 and 4 of this thesis, I develop theoretical models to analyze distinct aspects of firms' cyber-security incentives: Do they invest sufficiently in protection against cyber attacks? Given their security investment, do they collect too much data from the perspective of consumer privacy? And do they report cyber attacks promptly so that society can mitigate the damage? The strategic interactions between firms and consumers in this domain are characterised by externalities, moral hazard, and asymmetric information, implying the capacity for regulatory interventions to raise welfare. In the Chapter 3, I study the provision of insurance against ransomware attacks, and ask whether and under what conditions the presence of insurers can raise the revenue of hackers. With each one of my Chapters, I aim to directly contribute to active policy debates and inform the optimal design of policy.

Chapter 2, titled "Reputation and the Provision of Data Security", examines dig-

ital platforms' incentives to protect the vast amounts of consumer data they collect. Security investment is unobserved by consumers, and firms' incentives are reputational. I find that a planner can raise total consumer welfare by imposing ex-ante limits on data collection based on a firm's history of data breaches. Consumers may benefit from data collection, so collection limits affect them directly, but also indirectly by altering the firm's incentives for investment. The optimal policy qualitatively depends on whether firms or consumers control the level of data collection, and also on the extent to which consumers' privacy preferences are heterogeneous. Finally, the model is used to evaluate two policies akin to those already implemented in major jurisdictions: imposing minimum security standards and banning the retention of consumer data by firms.

Chapter 3, titled "The Market for Ransomware Insurance", develops a joint model of the market for ransomware insurance and the strategic interaction between firms and attackers. Hackers may or may not observe victims' insurance contracts, and firms may be restricted in paying ransom due to liquidity constraints. Insurance against business interruption has commitment value for firms in their bargaining with hackers and can reduce ransom demands. Under unobserved contracts, ransom demands and coverage are strategic complements, leading to multiple equilibria. Nevertheless, I find that in the absence of liquidity constraints, the presence of insurers benefits firms in every equilibrium. Imposing regulatory caps on insurance for ransom payments can select low-ransom equilibria when contracts are unobserved. Capping ransom coverage below firms' own liquidity ensures firms are made better off by the presence of insurers, and hackers worse off. The results of this Chapter inform the current debate on the social value of ransomware insurance markets, and provide clear recommendations for welfare-improving regulatory interventions.

Finally, prompt disclosure of cyber attacks is socially valuable, both for mitigation

of harms and also for information sharing among firms. However, firms have strong incentives to conceal incidents. This is the case for traditional data breaches and for ransomware attacks alike. Lenient regulatory treatment and support for firms that disclose breaches can encourage ex-post disclosure, but may sacrifice prevention incentives. This is the trade-off I study in Chapter 4, "Prevention and Disclosure of Data Breaches". In the model, a firm's disclosure decision is an optimal stopping problem; it only discloses a breach after becoming pessimistic about the prospect of privately concealing it. I characterize the optimal policy for a regulator who can commit to penalties following breach disclosure. If disclosure delay is ex-post verifiable and there is no other private information, a simple deadline policy can achieve the first-best outcome. If firms have private beliefs about concealability, the regulator uses delay-dependent penalties that induce immediate disclosure by pessimistic firms, but delayed disclosure by optimistic ones.

2 | Reputation and the Provision of Data Security

2.1 Introduction

In 2019, the phone numbers, Facebook IDs, names and locations of over 500 million Facebook users were accessed and sold on the dark web as a result of a security vulnerability [NPR, 2021]. Although such data breaches plausibly harm users who value their *privacy*, the frequency with which breaches occur raises suspicions about firms' incentives to protect the large amounts of personal data that they collect.

Digital platforms' incentives to invest in data security stem primarily from reputational concerns, because consumers cannot observe the level of data-security investments. Firms maintain a significant degree of ex-post discretion on how to implement their stated data-security policies: consumers cannot monitor how often software is patched, how much firms spend on data-security staff, or how diligently employees adhere to best practices. Anecdotal evidence corroborates the key role of reputation in the context of data security,[1][2] and the empirical work by Kamiya et al. [2021b] shows that following data breaches, firms suffer reputational damage beyond the directly-incurred penalties and litigation costs.

[1] According to the expert who uncovered the 2019 Facebook incident: "Individuals signing up to a *reputable* company like Facebook are trusting them with their data, and Facebook [is] supposed to treat the data with utmost respect…", see MIT Technology Review [2021]. Consumers need to rely on trust and the firm's reputation precisely because of the moral hazard problem detailed above.

[2] Discussing the reputational implications of data breaches, Caldwell [2016] notes that "The direct revenue losses … can be nearly negligible compared to the reputational damage incurred which in turn can lead to future revenue losses…"

However, despite the above, most models in the data protection literature (e.g. Fainmesser et al. [2023], Ahnert et al. [2022b], de Cornière and Taylor [2024]) are static and do not incorporate firm reputation in the analysis of firms' incentives. In this paper, I present a dynamic model of reputational concerns applied to the context of data security, in which investment is unobserved and there is uncertainty about the prevailing level of data-security risk. Such uncertainty is another prominent feature of this setting. For example, at any point in time, neither the firm nor the consumers know the exact rate at which data-breach attempts occur, or, as Gandal et al. [2022] emphasize, the extent to which available security measures will be effective protection against state-of-the-art adversarial methods.

Accounting for reputation and uncertainty allows me to derive novel implications of policies recently implemented by regulators in the EU, US, China and other jurisdictions. For example, the EU General Data Protection Regulation[3] (EU GDPR), has introduced notification requirements and penalties for data breaches (Articles 33 and 83), and has established the rights of consumers to decide how much of their data firms can collect (Article 7) and how long they can store it for (Article 17). Examined through my model, these existing policies affect firms' reputational incentives to invest in data security and thus the frequency of breaches and the extent to which society *learns* about cyber-security risk.

My model also allows me to analyse a *novel* policy of inherently dynamic nature: ex-ante limits on how much data can be collected by firms based on their history of data breaches. I find that such limits can indeed induce first-order improvements in consumer welfare. This is the case whether (a) firms choose how much data to collect from active users in each period or (b) users can choose precisely how much data to share with the firm in each period.

In the model, consumers with privacy concerns interact with a single digital plat-

[3]For examples of related legislation in other jurisdictions, see the California Consumer Privacy Act in the US or the Personal Information Protection Law of the People's Republic of China.

form over two periods. The firm does not charge for its service but requires the collection of personal data, which it monetises. Initially, I treat this level of required data collection as an exogenously determined feature of the service. To prevent data breaches, which harm consumers, the firm may in each period invest in data security, the level of which consumers do not observe. The prevailing level of security *risk* faced by the firm also influences the probability of a data breach at any level of the firm's investment, and both firm and consumers are uncertain about this risk. In this context, a firm's *reputation* is the perceived probability with which it faces low data risk. Bayesian consumers update their beliefs about risk based on whether or not a breach occurs in the first period and then decide on their second-period participation. Consumer retention motivates the firm to invest in the first period in order to avoid data breaches and the resulting harm to its reputation. In contrast, there are no investment incentives in the final (second) period of the model. An equilibrium in pure strategies always exists.

Before turning onto the evaluation of regulatory policies, I study how consumer surplus, CS, depends on equilibrium first-period investment. On the one hand, consumers benefit from fewer breaches in the first period. On the other hand, investment influences the informativeness of posterior beliefs about the underlying data risk. Whether the *learning effect* is positive or negative depends on the security technology in place,[4] and in particular on whether the data-breach probabilities under high and low risk converge at higher levels of investment.

Using the baseline model of fixed data collection, I assess common policies implemented in the EU and other jurisdictions: notification requirements (Lemma 3), minimum security standards (Proposition 4) and the imposition of consumer property rights in data (Propositions 5 and 6). To analyse the latter, I consider the effects of a relaxation of consumers' property rights that would allow the firm to retain their data between periods. Such data retention means that for first-

[4]The map from investment and risk level to the probability of a breach.

period users, part of the second-period expected privacy cost is *sunk*, and this causes a reduction in equilibrium investment. Thus, policies akin to the GDPR "right to erasure" will increase equilibrium investment.

Beyond these established policy levers, I extend the model to study a novel regulatory question: whether regulatory *limits* on data collection allow a planner to raise expected consumer surplus, and in particular limits that depend on whether the firm previously suffered a data breach or not. A key insight is that apart from directly affecting consumers' utility in the second period, such data caps will also affect firms' reputational incentives to invest in security.

I study this policy under two *regimes* of endogenous data collection: in the *firm regime*, the firm demands the profit-maximizing amount of data in every period. Instead, in the *consumer regime*, inspired by the "opt-out" regulation of the EU GDPR, active consumers have the option to share their preferred amount of data with the firm.

A main result of this paper is that in the firm regime, a regulator can increase total CS by (marginally) restricting data collection in the second period, regardless of whether the firm suffered a breach in the first period or not (Proposition 8). This result relies on two arguments: first, in the equilibrium of this regime, a profit-maximizing firm asks for so much data that consumers benefit from limiting data collection. Second, I show that marginal restrictions relative to equilibrium data collection have no first-order impact on the firm's investment incentives, hence only the above positive effect of limits remains.

I show how heterogeneity in privacy preferences can potentially reverse the result, in the natural case where the marginal consumer is also the most averse to data-sharing. I show that if the firm's revenue depends strongly on *participation* rather than data volume, it collects *too little* data from a consumer surplus perspective and the above policy prescription is reversed.

Even under the consumer regime, when consent policies are in place, there is scope for restricting data collection in ways that increase total consumer surplus. However, the policy recommendations are qualitatively different to those in the firm regime. When consumers benefit from higher equilibrium investment in the first-period,[5] I find that the planner should ex-ante restrict **only** the amount of data sharing with a firm that has suffered a breach in the first-period (Proposition 9). In *contrast* to the firm regime, the result is driven by the positive *indirect* effect of this policy on investment incentives.

In the following section, I review related literature. In Section 3, I present the baseline model of reputational concerns, derive the equilibrium and perform comparative statics. In Section 4, I analyse the impact of first-period investment on consumer surplus. Section 5 contains the analysis of policy counterfactuals that closely relate to existing policies. In Section 6, I introduce the two regimes of endogenous data collection and I analyse the effects of limits to data collection in each of the two regimes. Section 7 concludes. Unless found in the main text, all proofs are in the Appendix.

2.2 Related Literature

Methodologically, my model of reputational incentives relates to the seminal model of Kreps and Wilson [1982] and the career-concerns model of Holmström [1999] (as well as Benabou and Laroque [1992], Diamond [1989] and Mailath and Samuelson [2001], and Avery and Meyer [2012]), in which the agent's actions are motivated by incentives to influence the market's belief about a payoff-relevant characteristic.

Within the literature on data protection and privacy, the papers closest to mine are Toh [2018] and Jullien et al. [2020] and both feature unobserved actions by the firm in two-period models. The first paper also models firm reputation, but does

[5] A sufficient condition for this is that the learning effect of investment is positive.

not deal with data collection and learning considerations.[6] It focuses on optimal fraud liability between a bank and a retail website, on which transactions can potentially expose users' financial data. The website invests insufficiently in cyber security because investment is unobserved, and also the bank compensates the users for part of the harm. Similarly to my paper, the work of Jullien et al. [2020] is motivated by policy questions related to the EU GDPR, and features a model in which the firm's unobserved choice of how much consumer data to sell affect consumers' learning about their own vulnerability to having their data exploited. Unlike their work, my model features data protection and (transparent) data collection, thus our models are suited to study different policy counterfactuals.

The papers of de Cornière and Taylor [2024], Ahnert et al. [2022b], and Fainmesser et al. [2023] also study firms' incentives to invest in data security, and I contribute to this literature with a dynamic model that features reputational concerns. The key difference with my work is that they model the incentives of adversaries and endogenize the frequency of data-breach *attempts*. The presence of strategic adversaries introduces a negative network externality between users, since a large user base attracts more data-stealing attempts. de Cornière and Taylor [2024] study the interaction between firms' business models in duopoly and equilibrium levels of cyber security. Ahnert et al. [2022b] endogenize the hackers' choice between asking the firm for ransom or engaging in conventional data breaches and stealing users' data.[7] The work of Fainmesser et al. [2023] studies how a firm's business model determines incentives to store, monetize, and protect data. Firms that are more data-driven both collect more data and also invest more in data security. This complementarity arises because higher collection attracts more attackers and thus raises the marginal benefit of protection.[8]

[6] Our baseline models of reputation differ, too. There is no state uncertainty in Toh [2018], i.e., it does not relate to the models I mention at the beginning of this Section.

[7] I focus on "conventional" breaches in their language.

[8] In my model, an insight of similar nature leads to the convexity result in Proposition 2, and also to the comparative statics result in Lemma 1.

My work also contributes to the literature that studies firms' data collection and its regulation. The paper by Lefouili et al. [2024] is motivated by regulation that requires informed consent for data processing. The authors examine how caps on data monetization will affect firm incentives for (observable) investment in quality, which relates to my analysis in Section 6, but do so in a static model without consumer learning. The papers by Markovich and Yehezkel [2021] and Dosis and Sand-Zantman [2023] draw welfare comparisons between regimes of consumer- and firm-control of data collection in static models, also motivated by consent regulation, but neither studies data protection. The recent work by Ichihashi [2023] is related to my analysis of data retention, although in a model without data-security investments. In his work, the disutility of sharing additional data with the firm is lower for consumers whose type the firm can already accurately estimate.

Because of this, the platform can extract all information from (forward-looking) users in the long run, even if users would not be willing to immediately share that information in the first period.

Finally, the empirical work of Kamiya et al. [2021b] relates closely to my work. The authors use data on disclosed breaches for a sample of publicly listed firms in the US and find that they suffer *reputational* damage following disclosure of a data breach. The authors estimate that in the aftermath of a breach, out-of-pocket costs to the firm (e.g. litigation costs, penalties, consumer compensation) only account for a fraction of the drop in stock value. Consistently with my model, they argue that in a full-information world where disclosed cyber attacks reveal no information about either the firm or the state, the firm's loss of value should only reflect out-of-pocket losses. Thus, their paper provides valuable empirical justification for my model of reputational incentives.[9]

[9]Relatedly, Peukert et al. [2022] find that *lower traffic* websites made fewer calls to third party vendors (i.e., used less consumer data) following the GDPR's implementation. An explanation consistent with consumers being responsive to firms' reputations in their data-sharing decisions

More generally, there are excellent surveys on the economics of privacy and data protection, both very recent, by Goldfarb and Tucker [2023], as well as slightly older, by Acquisti et al. [2016]. The former also covers recent empirical work, both on the economic impact of the GDPR and on measuring privacy concerns. The latter deals in depth with the theoretical literature on the economics of privacy.

2.3 Baseline model

2.3.1 Consumers

A single firm provides a digital service and interacts with a unit mass of potential consumers over $T = 2$ periods. At the beginning of each period, consumers decide whether to use the service. The firm does not charge monetary registration or usage fees, but requires that users share their personal data with it.

Each consumer has type θ, i.i.d drawn from distribution G on $[0, 1]$. Users derive utility $v(d) - \theta$ from using the service, where $d \in [d^{min}, d^{max}]$ is the amount of *data* that they share with the firm, and $v(d), v'(d) > 0$. However, users also suffer harm $L(d)$ in the event of a *data breach* and $L'(d) > 0$, so that breaches are more harmful when the firm collects more data.[10] In this section I will treat d as exogenous, and constant across periods; I later endogenize data in Section 2.6. The above implies that for given probability of a breach, p, the expected utility of a user is:

$$U(d, p, \theta) := v(d) - pL(d) - \theta \tag{2.1}$$

is that consumers have less information about prior breaches of lower-traffic websites and thus choose to opt out of inessential data collection. The authors do not explicitly interpret this finding.

[10] I do not distinguish between taste-based and instrumental preferences for privacy. Lin [2022] experimentally separates these two sources of value for privacy and finds strong evidence for the existence of both.

Users that have suffered a breach in the first period will incur additional loss of $L(d)$ if they are active again in the second and the firm suffers an additional breach. The disutility of an active user that experiences a breach in the second period is independent of both their first-period participation and of whether a breach previously occurred or not.[11] Hence, consumers only consider current-period expected payoffs when making participation decisions and the indifferent type, $\theta(d,p)$ satisfies $U(d,p,\theta(d,p)) = 0 \iff \theta(d,p) = v(d) - pL(d)$, with $\partial\theta(d,p)/\partial p = -L(d) < 0$. The mass of active users in a given period is $G(\theta(d,p))$. For convenience of exposition, I focus on parameter cases such that there is always an indifferent type in the interior of $[0,1]$.

2.3.2 Investment and data breaches

In each period, the firm can invest in costly data security to decrease the probability of a data breach. This investment can be thought of as expenditure for hiring IT staff and security specialists, investing in state-of-the-art software and hardware, or adopting work processes that enhance data security at the expense of productivity, e.g. multi-factor authentication.[12] It can also be thought of as effort that firms exert to better screen potentially malicious third-party web service vendors that get access to consumer data. I denote investment by e. Importantly, I assume that this investment is ex-ante and ex-post *unobserved* by the consumers. Even if a data breach is made publicly known, it could be quite costly, if at all feasible, to understand the extent to which it occurred due to lack of due diligence by the firm.

I denote the "outcome" of each period by the random variable y_t, which can take values $\{b,n\}$, standing for *breach* and *no breach*. The outcome becomes publicly known at the end of each period and $f^r(e) := P(y_t = b|r,e_t)$ is the

[11] I relax the first assumption in the section on data retention.

[12] The empirical work of Hastings et al. [2023] measures the cost of more stringent authentication processes for an organization.

probability a *breach* occurs, given investment e by the firm in that period and given the firm's type, $r \in \{h, \ell\}$, standing for high- and low-risk, respectively. I will refer to the pair of functions $\{f^\ell, f^h\}$ as the firm's *technology*. For either type r, a breach is less likely at higher levels of investment, i.e., $\partial f^r(e)/\partial e < 0$. For convenience of exposition, I will focus on technologies that are differentiable in e, for all $e \in [0, 1]$.[13]

The following assumption defines the low-risk type $r = \ell$ in this model:

Assumption 1: $f^h(e) \geq f^\ell(e)$ for all $e \in [0, 1]$, and $f^h(0) > f^\ell(0)$.

A breach is weakly less likely for the *low*-risk type, and strictly so when $e = 0$; the significance of the second part will become clearer in the equilibrium derivation section. I assume that neither the firm nor the consumers know the value of r, and I will refer to the firm's *reputation*, as the probability with which consumers believe that the firm's type is ℓ in a given information set of the extensive form game. The firm has prior reputation $P(\ell) = \mu \in (0, 1)$, which is common knowledge. After observing the outcome at the end of $t = 1$, consumers update their beliefs about r, given the investment level that they conjecture the firm used in the first period. Under Assumption 1 and for $\mu \in (0, 1)$, "no breach" occurs with positive probability, and the posterior belief following that outcome, as a function of first-period investment e_1, is found via Bayes' Rule:

$$\mu_n(e_1) := P(r = \ell | y_1 = n, e_1) = \frac{\mu(1 - f^\ell(e_1))}{\mu(1 - f^\ell(e_1)) + (1 - \mu)(1 - f^h(e_1))} \tag{2.2}$$

I similarly define $\mu_b(e_1) := P(r = \ell | y_1 = b, e_1)$ as the posterior following a breach in the first period. When both posteriors are well defined, $\mu_n(e_1) > \mu > \mu_b(e_1) \iff f^h(e_1) > f^\ell(e_1)$. Intuitively, lack of a breach is evidence of type $r = \ell$ if that type has lower probability of suffering a breach at investment e_1.

[13] This assumption excludes kinked cases in which there exists $\hat{e}$ strictly in the interior of $[0, 1]$ such that $f^\ell(e) = 0 \iff e > \hat{e}$. Treating such cases offers no additional insight.

Regarding notation, throughout this paper I make use of the subscript $s \in \{1, n, b\}$ to denote the *subgame* of the extensive form that variables refer to. I use $s = 1$ for variables that refer to the first period of the game, and $s = b$ (respectively, $s = n$) for those that refer to the second-period subgame following a $y_1 = b$ ($y_1 = n$) realization.

2.3.3 Reputational incentives and equilibrium

I now turn to the firm's profit-maximization problem. 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., $r'(d) > 0$. Total revenue in a given period as a function of data collection and the expected breach probability is $\Pi(d, p) := r(d)G(\theta(d, p))$, where $\theta(d, p)$ is the indifferent consumer type. Investment in the first period is unobserved and consumers base their decisions on their *conjecture* $\tilde{e_1}$, which the firm cannot influence through its choice of investment. For given conjecture and reputation, consumers anticipate a breach to occur with probability $p(\mu, \tilde{e}) := \mathbb{E}_\mu[f^r(\tilde{e})] = \mu f^\ell(\tilde{e}) + (1-\mu) f^h(\tilde{e})$. Accordingly, the conjecture determines posterior reputations $\mu_n(\tilde{e_1}), \mu_b(\tilde{e_1})$ which the firm also cannot influence.[14] First-period investment will thus *not* affect first-period revenue, and will only affect expected total profit via the (actual) probability of a breach, $p(\mu, e)$, and the cost of investment, $C(e)$. I assume the latter to be increasing, convex, and satisfying $C(0) = C'(0) = 0$. Assuming no discounting, for simplicity, total expected profit is:

$$\begin{aligned}\mathbb{E}\Pi(d, e_1; \tilde{e_1}) = \Pi(d, \tilde{p_1}) - C(e_1) \\ + p(\mu, e_1)\Pi(d, \tilde{p_b}) + (1 - p(\mu, e_1))\Pi(d, \tilde{p_n})\end{aligned} \tag{2.3}$$

[14] Even though the conjecture will have to be correct in equilibrium.

I use the shorthand notation $\tilde{p_n} = p(\mu_n(\tilde{e_1}), 0)$ and $\tilde{p_b} = p(\mu_b(\tilde{e_1}), 0)$, to denote posterior expected breach probabilities that guide consumers' decisions in period 2.[15] The second argument is zero because the firm does not invest in any information set of period 2. It bears no direct loss in the event of a breach, and there is no user retention incentive in the last period of the game, so that in any equilibrium, $e_n = e_b = 0$.

Table 2.1: Notation: Beliefs and Technology

Symbol	**Definition**
e, $\tilde{e}$	True and conjectured investment
$r \in \{h, \ell\}$	Firm's risk type (high or low)
$f^r(e)$	Probability of a data breach for type r and investment e
μ	Prior belief that $r = l$
μ_n, μ_b	Posterior beliefs after no breach and breach, respectively
p_1, p_n, p_b	True breach probabilities
$\tilde{p_1}, \tilde{p_n}, \tilde{p_b}$	Conjectured breach probabilities

I emphasize that the firm can influence the probability with which it suffers a breach, i.e.,whether the perceived posterior probability is $\tilde{p_n}$ or $\tilde{p_b}$, but not the values of $\tilde{p_n}, \tilde{p_b}$. In the first period, the firm chooses e_1 to maximize $\mathbb{E}\Pi(d, e_1; \tilde{e_1})$, and the first-order condition that must be satisfied at an interior solution is:

$$MB(e_1; \tilde{e}_1) := \overbrace{-\frac{\partial p(\mu, e_1)}{\partial e_1}}^{(+)} \underbrace{\Big(\Pi(d, \tilde{p_n}) - \Pi(d, \tilde{p_b})\Big)}_{\text{Reputational Premium}} = C'(e_1) \tag{2.4}$$

According to equation (2.4), the monopolist's best response is to invest up to the point where marginal cost equals the marginal benefit of investment *given* con-

[15] More generally, when it is clear, I suppress dependence on e_1 and write μ_s instead of $\mu_s(e_1)$, for $s \in \{n, b\}$.

sumers' conjectures, $MB(e_1; \tilde{e}_1)$. Since total revenue Π is increasing in participation and, thus, decreasing in the posterior probability of a breach, the reputational premium is positive iff $\tilde{p_n} < \tilde{p_b}$. This is the case for all $\tilde{e_1}$, under Assumption 1, according to which a low-risk type suffers a breach with weakly lower probability for all e, hence $\tilde{p_n} < \tilde{p_b}$.[16] Greater difference between revenue in the two potential outcomes induces higher investment provision, and so does greater (ex-ante expected) efficiency of investment in reducing the probability of a breach. To turn (2.4) into an equilibrium-defining equation, I must impose the equilibrium condition that conjectures are correct i.e., $\tilde{e_1} = e_1$. In the following Proposition, I state that a stable Perfect Bayesian Equilibrium in pure strategies always exists. In equilibrium, the firm best responds to consumer beliefs in every information set, which implies $e_n^* = e_b^* = 0$, and consumers anticipate the correct investment levels and make optimal participation decisions. Beliefs are updated according to Bayes' rule and to simplify matters, I assume that $f^h(1) > 0$ which guarantees both first-period outcomes are realized with positive probability for any e.[17]

Proposition 1. *A pure-strategy Perfect Bayesian Equilibrium always exists. For given parameter values, define e_1^* to be the smallest first-period investment value across (potentially multiple) pure-strategy equilibria. e_1^* is always strictly positive and there are two possible cases:*

- *$e_1^* < 1$ and e_1^* satisfies the equilibrium first-order condition.*
- *There is a unique equilibrium with $e_1^* = 1$. This case can only occur if*

[16] In fact, the second part of Assumption 1 is the only one necessary for the equilibrium derivation argument, and the only one I make use of in the proof. An investment level such that $f^h(e) < f^\ell(e)$ could never be part of an equilibrium. Looking at expression (2.2), this would imply $\mu_n(e) < \mu$, meaning "no-breach" would be evidence of a **high**-risk type and the reputation premium would be negative. However, the additional generality does not offer useful insight, and for clearer exposition, I make use of the stricter assumption $f^h(e) \geq f^\ell(e)$, for all e.

[17] This means that the high-risk type always suffers a breach with positive probability, and combined with our existing assumption strictly interior μ, we obtain the statement, since the low-risk type avoids a breach with positive probability, for any e, by Assumption 1. This assumption is well in line with the reality that data breaches can never be fully prevented, not least because of human error.

$f^h(e) > f^\ell(e)$ *for all* $e \in [0,1]$ *and* $C'(1)$ *is sufficiently low.*

For sufficiently convex cost of investment, equilibrium is unique.

From now on, I will use "equilibrium" to refer to the Perfect Bayesian Equilibrium that features the smallest first-period investment level. Notice that $e_1 = 0$ can never be part of an equilibrium under the assumption of $C'(0) = 0$: by the second part of Assumption 1, $\mu_n(0) > \mu > \mu_b(0)$, which implies that the reputational premium from avoiding a breach at that investment level is positive. I derive the following comparative statics result for the equilibrium value e_1^*.

Lemma 1. *For* $\theta \sim U$, e_1^* *is increasing in* d.

There are two economic forces behind this result: first, $r'(d) > 0$, i.e., revenue per consumer increases in data collection, which increases the reputational premium holding posterior beliefs of consumers fixed. Second, $\frac{\partial^2 \theta(d,p)}{\partial d \partial p} < 0$, i.e., the mass of active users is more sensitive to the probability of a breach (i.e., reputation) at higher levels of d: in turn, this is because I assume the disutility induced to consumers from a breach, $L(d)$, is *increasing* in data collection. Both effects suggest $\frac{\partial^2 \Pi}{\partial d \partial p} < 0$, which leads to the above result. Thus, higher data collection across the two subgames of period 2 acts as a *commitment device* in this model.

It is not obvious whether higher prior probability μ will induce higher e_1 in equilibrium, since both posterior breach probabilities p_n, p_b are decreasing in the prior, holding investment e_1 fixed. Similar to the model of Benabou and Laroque [1992] in which investment is also purely motivated by learning concerns, quasi-concavity of $e_1^*(\mu)$ obtains: reputational incentives are maximized at intermediate levels of prior beliefs at which uncertainty about the firm's type is the greatest. Intuitively, a firm with $\mu \sim 0$ will not convince users it is of low-risk even if it avoids a breach in the first period. I illustrate this in Figure 2.1 and prove the following Lemma:

Lemma 2. *If* $f^\ell(e) > 0$, *for all* $e \in [0,1]$ *and if* $\theta \sim U$, e_1^* *is strictly quasi-concave*

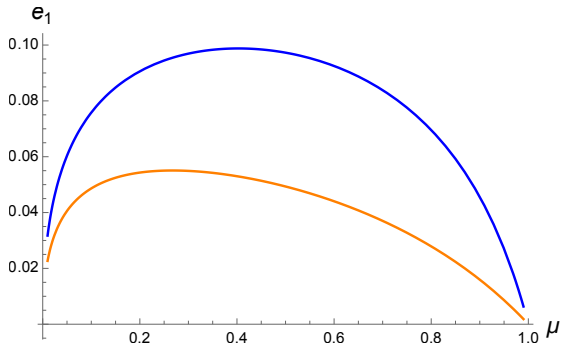


Figure 2.1: The two different curves correspond to the following specifications for the technology, f. Blue: $\{f^\ell(e) = 0.2, f^h(e) = 1 - 0.8e\}$. Orange: $\{f^\ell(e) = 0.4, f^h(e) = 1 - 0.6e\}$. Drawn for linear $r(d)$ and for $U(d, p, \theta)$ quadratic-concave in d.

in μ, and approaches zero as $\mu \to 1$ or $\mu \to 0$.

We can extend the baseline model to account for imperfect observability of data-breaches. We do so in a reduced form way, via introducing a parameter $q \in [0, 1]$, which stands for the *exogenous* probability that a breach becomes public information after it has occurred. 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 a breach is observed is $p(\mu, e)q$, increasing in q and decreasing in e. I assume that the same observability probability q applies to both types $r \in \{\ell, h\}$, so the posterior μ_b is unaffected by q: $\mu_b = q\mu f^\ell(e)/(qp(\mu, e)) = \mu f^\ell(e)/p(\mu, e)$. On the other hand, the posterior μ_n is always *increasing* in q, since at higher observability rates, no observation of a breach becomes more informative about whether a breach occurred or not and thus about the firm's type. The level of q has an additional *direct*, positive effect on investment incentives, and we obtain the following:

Lemma 3. *The equilibrium level of investment e_1^* is increasing in the probability of observing a breach, q.*

This result naturally relates to policies which require firms that suffer data breaches to notify the regulator and their consumers, see for instance the recent amendment adopted by the US Securities and Exchange Commission (SEC) in May 2024

(Parks [2024]),[18] and suggests that apart from any direct benefits of consumer notification, there are equilibrium effects on investment incentives. By raising e_1^*, a policy that increases q will also increase first-period equilibrium participation.

2.3.4 Discussion of assumptions

Before moving on to the welfare and policy analysis, I discuss two of the main assumptions underlying this model.

Implicit incentives: In this model, a data breach is not directly damaging to the firm, or at least the direct harm is relatively minor compared to the reputation damage, consistent with the findings of Kamiya et al. [2021b]. As Koutroumpis et al. [2022] mention, data breaches (which are a *subset* of all cyber attacks) are often harmless to the victim firm in terms of operations disruption, which is also the case for the Facebook examples in my Introduction.

Such data breaches are also the motivating application of my paper. An alternative is *ransomware attacks*, which are directly costly to the firms that have to pay ransom to restore their digital operations and retrieve stolen data. These are also potentially *harmless* to consumers, if hackers do not exploit stolen data after firms pay the ransom.[19]

Symmetric information: In this model, firms and consumers are symmetrically (un)informed about the firm's type, r. Firms and consumers both understand that more investment leads on average to fewer breaches, but are uncertain of the exact relationship. In reality, firms (and consumers) do, indeed, face substantial uncertainty about the efficacy of data-protection measures against data breaches and also about the methods available to malicious parties. The term *zero-day*

[18]"..requiring SEC-regulated investment advisers, investment companies, and broker dealers to notify individuals whose sensitive customer information was, or is reasonably likely to have been, accessed or used without authorization within 30 days of becoming aware...".

[19]In other words, ransomware attacks and typical data-breaches differ in the *distribution of losses* among firms and consumers. This is the starting point of Ahnert et al. [2022b], who endogenize hackers' decision to choose between these models of operation.

vulnerability precisely captures such unaddressed security flaws. This uncertainty is the subject of empirical studies by Gandal et al. [2022] and Gandal et al. [2020] who note that: "... little if anything is known about the relationship" among vulnerabilities, preventive measures, and security incidents, like the leaking of sensitive data to the web. My modeling assumption is that firms have effectively no additional information relative to consumers.

2.4 Consumer Surplus

Before analysing policy changes and their impact on equilibrium investment by the firm, I examine how security investment affects consumer surplus. In this section, I look at how changes in first-period investment, e_1, affect expected consumer surplus in the first and second period of the game. Consider, for instance, changes in revenue per consumer, or in the marginal cost of investment, or increases in a monetary penalty levied on the firm if a breach is observed in the first period. Each of these changes will not affect consumer surplus directly, but will do so indirectly by changing equilibrium investment in the first period, e_1^*. I define *equilibrium* first-period expected consumer surplus as a function of first-period investment and data collection. This is aggregate expected utility over consumers that are active when the probability of a breach is $p_1(e_1) = p(\mu, e_1)$:

$$CS_1(e_1, d) := \int_0^{\theta(d, p_1(e_1))} U\big(d, p_1(e_1), \theta\big)\, dG(\theta) \tag{2.5}$$

Higher investment causes a decrease in the expected number of breaches and an increase in expected utility for inframarginal consumers. It will also cause an increase in the mass of first-period active users. Furthermore, as security investment increases, each additional marginal reduction in the probability of a breach applies to an increasingly large user base and, thus, contributes increasingly more to aggregate consumer surplus. I formalize these arguments in the next Proposition,

in which I also make use of the following definition:

Definition 1: A *linear* technology is a collection of weakly positive scalars $\{\ell_0, \ell_1, h_0, h_1\}$, such that breach probabilities for the two types are $f^\ell(e) = \ell_0 - \ell_1 e$ and $f^h(e) = h_0 - h_1 e$, and $f^h(e), f^\ell(e)$ take values in $[0,1]$ for all $e \in [0,1]$.

Proposition 2. *For any $d \in [d^{min}, d^{max}]$, first-period expected consumer surplus, $CS_1(d, e_1)$, is increasing in first-period investment, e_1. In addition, if the technology is linear, $CS_1(d, e_1)$ is also convex in e_1.*

Similarly, I define *equilibrium* second-period (ex-ante) expected consumer surplus, which depends on first-period investment through its effect on *learning* about the firm's type. The fact that in equilibrium $e_n = e_b = 0$ implies that, in subgame $s \in \{n, b\}$, a breach occurs with probability $p_s(e_1) = p(\mu_s(e_1), 0)$, and I define:

$$CS_2(e_1, d) := E_{\mu, e_1}\left[\int_0^{\theta(d,\, p_s(e_1))} U\big(d, p_s(e_1), \theta\big)\, dG(\theta)\right] \tag{2.6}$$

The expectation is taken over the distribution of types and the distribution of posterior beliefs, given e_1. First-period investment influences whether a breach occurs or not in that period and thus the distribution of posterior beliefs about the firm's risk type. Given their posterior beliefs in each information set, consumers make optimal participation decisions. But whether those optimal decisions better match the firm's true type in equilibria with higher levels of e_1 depends on first-period outcomes are more informative in such equilibria. The following definitions are helpful:

Definition 2: If the technology is such that $\mu_n'(e_1) \geq 0$ and $\mu_b'(e_1) \leq 0$ for all $e \in (0,1)$, I will say that investment *enables* learning.

Definition 3: If the technology is such that $\mu_n'(e_1) \leq 0$ and $\mu_b'(e_1) \geq 0$ for all $e \in (0,1)$, I will say that investment *impedes* learning.[20]

[20]Equivalently, the distribution of posterior beliefs $(\mu_n(e_1), \mu_b(e_1); p(\mu_1, e_1))$ is indexed by e_1. When e_1 impedes learning, distributions of posteriors with lower e_1 dominate those with higher

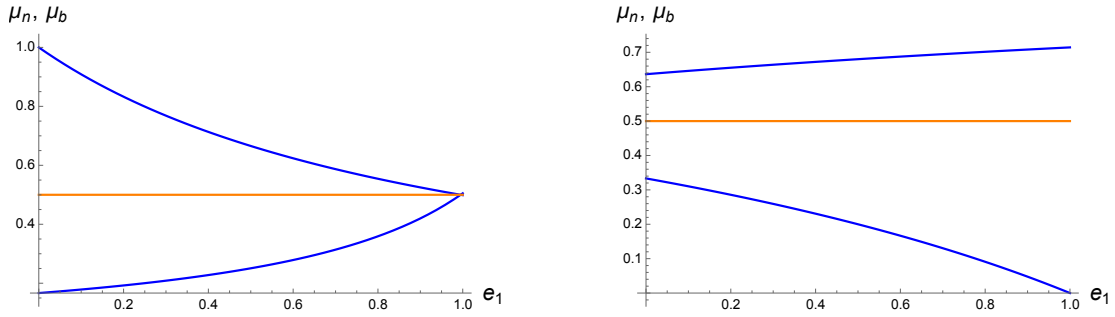


Figure 2.2: Illustration of Definitions 2 and 3 for a linear technology f (Definition 1). Left: Example technology that satisfies Definition 2 $\{f^\ell(e) = 0.2, f^h(e) = 1 - 0.8e\}$. Right: Example technology that satisfies Definition 1 $\{f^\ell(e) = 0.3 - 0.3e, f^h(e) = 0.6\}$. Drawn for $\mu = 0.5$, shown at the orange horizontal line.

If $\mu_n'(e_1) > 0$ and $\mu_b'(e_1) < 0$, then at higher first-period investment a breach becomes stronger evidence of a high-risk type and lack of a breach is stronger evidence of a low-risk type. On the other hand, if $\mu_n'(e_1) < 0$ and $\mu_b'(e_1) > 0$, the two posteriors converge towards the prior as first-period investment increases.[21] Given these definitions, we state the following result:

Proposition 3. *For any $d \in [d^{min}, d^{max}]$:*

(a) If the technology is such that investment impedes learning, ex-ante expected second-period consumer surplus $CS_2(e_1, d)$ is decreasing in investment e_1. If the technology is such that investment enables learning, then $CS_2(e_1, d)$ is increasing in investment e_1.

(b) If investment either impedes or enables learning, and technology is linear, then $CS_2(e_1, d)$ is convex in e_1.

Part (a) of the Proposition confirms that when consumers' posterior beliefs become less informative about the firm's type, they can make less appropriate decisions in period 2 and expected consumer surplus of that period decreases. To understand

e_1 in the convex order, for all $e_1 \in [0, 1]$.

[21] It should be intuitive that if investment impedes learning, the reputation premium is decreasing in consumers' conjecture about investment. At higher levels of conjectured investment, the gap between the two posterior beliefs shrinks and the benefit to achieving high reputation decreases. This is also why a equilibrium uniqueness is guaranteed in this case. The opposite holds if investment enables learning.

the convexity result of part (b), some additional explanation of the results in part (a) is helpful. First-period investment affects equilibrium second-period expected utility via two channels: First, it affects how frequently each type achieves high reputation in period 2, holding posterior reputations fixed. Second, it affects the firm's posterior reputation in each information set of period two. Posterior reputations only affect consumer surplus via affecting consumer decisions, but since these decisions are made by consumers optimally, there are no first-order effects of investment via the posterior beliefs channel. Thus, the only first-order effect is via the frequency with which each type avoids a breach, holding consumers' posterior beliefs (and thus decisions) fixed. Convexity obtains because regardless of the sign of this first-order effect, the second-order effect is always positive: as consumers adapt their optimal decisions, they either mitigate the negative impact of additional reductions in the posteriors' informativeness or, when e_1 enables learning, accentuate the positive impact of additional increases in information available to them in the second period.

Propositions 2 and 3 imply the following important Corollary:

Corollary 1. *If the technology is linear in e, total expected consumer surplus across both periods is convex in* e_1*. If investment enables learning, total consumer surplus is strictly increasing in first-period investment. If it impedes learning, total consumer surplus is potentially non-monotonic in first-period investment, and, for any set of parameter values, if it is decreasing at any* $\hat{e_1} > 0$ *is also decreasing at* $e_1 < \hat{e_1}$*.*

The above Corollary has implications for policy: the benefits to consumers from increases in equilibrium investments in security are larger in equilibria with high initial levels of investment in data security. Suppose a regulator wants to maximize consumer surplus and has the option of imposing monetary penalties to firms that suffer breaches at the end of the first-period. This regulator should be more willing

to increase the value of the penalty starting from equilibria with initially high value of investment, i.e., equilibria with relatively few breaches. Additionally, if the technology is such that investment impedes learning, the planner must be wary of levying penalties when investment is low in the initial equilibrium: numerical evidence in Figure 3 shows that consumer surplus can indeed be *decreasing* in first-period investment. To understand when this can arise, we must first understand when investment impedes learning. The following result answers this question.

Lemma 4. *Within the class of linear technologies that satisfy Assumption 1, investment impedes learning if and only if* $\ell_1 h_0 < h_1 \ell_0$ *and investment enables learning if and only if* $\ell_1(1-h_0) > h_1(1-\ell_0)$.

When the technology is linear and satisfies Assumption 1, it must be that $h_0 > \ell_0$. Thus, according to the above Proposition, for higher first-period investment to impede learning, it must be that investment is relatively more effective in reducing the probability of a breach when risk is high, i.e., that $\ell_1 < h_1$. On the other hand, a necessary (but not sufficient) condition for investment to enable learning is that the marginal efficiency of investment is higher when $r = \ell$.[22] It may be that the probability of a breach is lower for all e when $r = \ell$, but it still is necessary to consider the relative *marginal* efficiency of investment between $r = \ell$ and $r = h$ to understand whether it enables or impedes learning. Finally, whether investment enables or impedes learning depends only on the specification of the technology and not on the prior μ.

I illustrate the results of this section using Figure 3: The decreasing slope of total consumer surplus at $e_1 = 0$ obtains when (a) the productive (marginal) effect of fewer breaches is of low magnitude and (b) the negative learning effect is of high magnitude. Following the arguments behind Propositions 2 and 3, this is,

[22] This is consistent with findings in other models of reputation concerns, see the survey by Bar-Isaac and Tadelis [2008]. The "Imitation" and "Separation" cases they point at are special cases of the classes I identify.

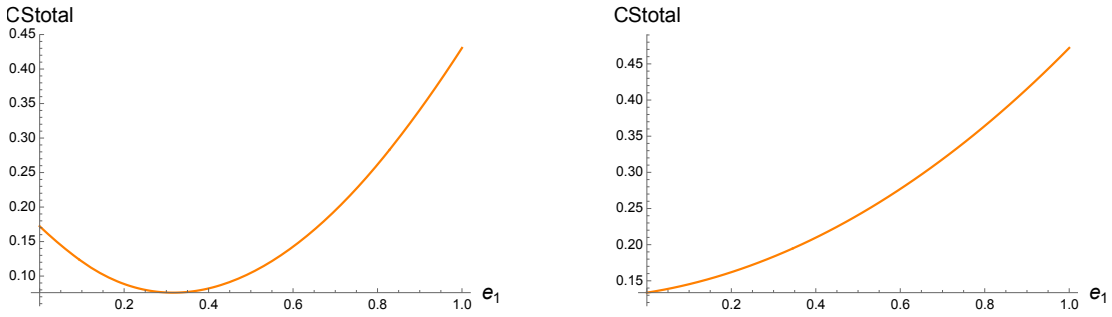


Figure 2.3: The vertical axis is the sum of ex-ante expected consumer surplus across both periods. It is always convex in e_1. Left: Illustration of Corollary 1, for the linear technology $\{f^\ell(e) = 0.2, f^h(e) = 1 - 0.8e\}$. Right: Illustration of Corollary 1, for the linear technology$\{f^\ell(e) = 0.3 - 0.3e, f^h(e) = 0.6\}$. The non-monotonicity can only appear when $h_1 > \ell_1$. Drawn for $U(d, p, \theta) = \alpha d - (1 + p)d^2 - \theta$ and $\mu = 0.5, d = 1.6, \alpha = 2.5$.

intuitively, the case when (a) at $e_1 = 0$ a small mass of agents participates in the first period, so that the effect of increased security only applies to a small mass of agents, and (b) there is a large difference between participation in the two information sets of the second period, meaning that first-period outcomes are very informative about risk; when this is true, the information generated by the first-period outcome has greater value for second-period consumer surplus. On the left panel of Figure 2.2, $h_0 = 1$ implies that good news are perfectly informative at $e_1 = 0$. The information structure is such that both good and bad news lead to large changes in beliefs, leading to a (negative) learning effect of high magnitude and the corresponding non-monotonicity in the left-panel of Figure 2.3.

Discussion. Before concluding this section, I discuss the interpretation of the condition in Lemma 4. The sign of $(h_1 - \ell_1)$ relates to our interpretation of the firm's risk type, r and we can consider different interpretations that are consistent with either sign: for example, if the risk type reflects the overall *intensity* of malicious activity, e.g. the exogenous rate of attempts to steal data from firms similar to the focal one, then investment is plausibly more useful in periods of heightened activity and innovation, i.e., $h_1 > \ell_1$ and investment does not enable learning.

An alternative interpretation of risk is that it reflects whether a substantial innovation has occurred with respect to the methods that malicious parties use to breach firms' networks. If the body of knowledge available in the community of security researchers has not yet incorporated these new methods, it is plausible that investments in security will have little effect on reducing the probability of a breach. If $r = \ell$ corresponds to no major hacking innovation having taken place, and $r = h$ corresponds to the opposite, then e is plausibly more efficient in the "pre-innovation" setting of $r = \ell$. Under this interpretation, $\ell_1 > h_1$ is the more reasonable assumption and, by Lemma 4, investment does not impede learning.

2.5 Policy Counterfactuals

Having understood how investment e_1 affects first- and second-period consumer surplus in this model, I examine policies that affect equilibrium investment.

2.5.1 Minimum security standards

Next, I analyse the implications of minimum security standards for a firm's reputational incentives. Suppose that a planner can enforce a minimum level of investment $\underline{e} \in (0, 1]$ in both periods, such that the firm now is restricted to choosing investment $e \in [\underline{e}, 1]$ in every information set.[23] In equilibrium, second-period investment in any information set will be $e_n^* = e_b^* = \underline{e}$. But how does this policy affect first-period investment, e_1^*? If at the initial equilibrium $e_1^* < \underline{e}$, then the new equilibrium investment will be at least as large as $\underline{e}$. Consider first the opposite case, and in particular the impact on e_1^* of an increase of minimum standards from an initial value of zero to $\underline{e} < e_1^*$. Holding consumer conjectures fixed at e_1^*, how does this change the reputational premium earned by the firm

[23] The cost and marginal cost functions take as argument the firm's total investment, i.e., the minimum imposed by the regulator and any additional voluntary investment by the firm. I am assuming that the firm's IR constraint is always satisfied.

in case it avoids a breach? The reputational premium relates to the difference in posterior breach probabilities between the two information sets of the second period. This difference is given by:

$$\begin{aligned} & p(\mu_b, \underline{e}) - p(\mu_n, \underline{e}) \\ =&[\mu_b f^\ell(\underline{e}) + (1-\mu_b) f^h(\underline{e})] - [\mu_n f^\ell(\underline{e}) + (1-\mu_n) f^h(\underline{e})] \\ =&(\mu_n - \mu_b)(f^h(\underline{e}) - f^\ell(\underline{e})) \\ =&(\mu_n - \mu_b)(h_0 - \ell_0) + (\mu_n - \mu_b)(\ell_1 - h_1)\underline{e} \end{aligned} \tag{2.7}$$

and the last equality assumes technology *linear* in e. Whether this difference will increase or decrease in $\underline{e}$ depends on the sign of $(\ell_1 - h_1)$, i.e., on whether investment has higher marginal productivity (in reducing the probability of a breach) for high- or low-risk firms. When investment offers higher marginal security for the high-risk type, imposing minimum standard $\underline{e}$ means that breach probabilities for the two types will converge in the second period of the game as a result of imposing minimum security standards; this reduces the firm's incentive to be perceived as a low type and, thus, its reputational incentive for investing in the first period.[24] The opposite result obtains when $\ell_1 > h_1$: in that case, the minimum security level $\underline{e}$ magnifies the difference between types in the second period and acts as a *commitment device*, complementing the firm's existing implicit incentives. I show the following:

Proposition 4. *Assume that technology is linear and $\theta \sim U$. Additionally, assume that cost is sufficiently convex so that there is a unique equilibrium.*
(a) If investment impedes learning: If for an initial value of $\underline{e}$, the equilibrium first-period investment satisfies $e_1^* > \underline{e}$, then a marginal increase in $\underline{e}$ will reduce e_1^*. If $e_1^* = \underline{e}$, then an increase in the minimum standard $\underline{e}$ will induce $e_1^* = \underline{e}$ in

[24] This argument is closely related to the intuition that leads to Lemma 4. I remind the reader that $\ell_1 > h_1$ is a necessary but not sufficient condition for e_1 to enable learning.

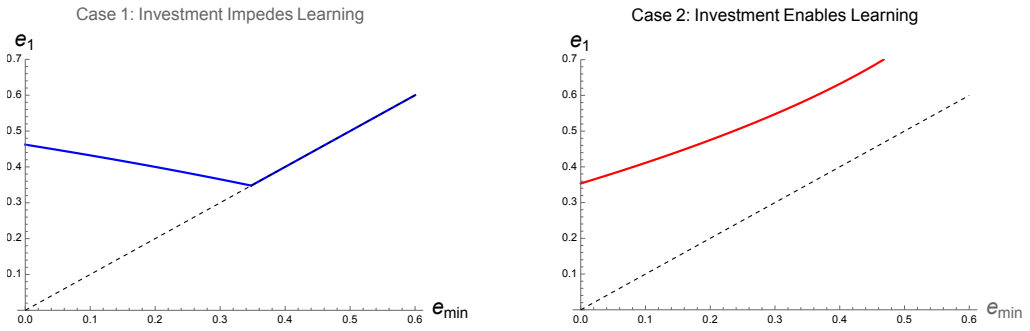


Figure 2.4: First-period equilibrium investment as a function of the minimum standard $e_{\min}$. Illustration of Proposition 4, for uniform distribution of consumer types. On the left (right), a technology is used such that investment to impedes (enables) learning. The dashed 45-degree line is drawn in both figures. When investment impedes learning, e_1^* is decreasing until the constraint becomes binding.

the new equilibrium.

(b) If investment enables learning: if for an initial value of $\underline{e}$, it holds that $\underline{e} < e_1^*$, a marginal increase in $\underline{e}$ will induce an increase in e_1^*. If $e_1^* = \underline{e}$, then an increase in the minimum standard $\underline{e}$ will induce $e_1^* \geq \underline{e}$ in the new equilibrium.

Thus, when $h_1 > \ell_1$, which is true when investment impedes learning, attempts to raise e_2 by imposing minimum security standards will erode the reputational premium and decrease e_1, if initially $e_1 > \underline{e}$, but this potential trade-off will not exist if $\ell_1 > h_1$. The discussion in the previous section regarding situations in which investment impedes/enables learning applies again.

2.5.2 Ban on data retention

In this section, I extend the model to analyse the impact of data retention on equilibrium incentives of the firm and consumers. By data retention, I refer to firms' practice of storing users' data even after those users have stopped using the service. I assume that users whose data is retained by the platform suffer harm from second-period data breaches, regardless of whether they use the service in that period or not. Extending the model in this direction will naturally allow to

assess the impact of introducing a "right to be forgotten", i.e., allowing consumers who no longer use a service to have their personal data deleted.[25]

I introduce parameter $\beta \in [0,1]$, which captures the degree to which the firm retains past consumers' data. I will focus on the case of $\theta \sim U$, so first-period participation will be given by an endogenous cutoff value $\theta_1 := \theta(d, p(\mu, e_1))$, and I similarly use the shorthand notation θ_n and θ_b. However, consumers' first-period decision now has implications for their second-period payoff, and consumers will make decisions to maximize their expected discounted total utility across periods. A user participates in a second-period information set if their expected utility from doing so exceeds their outside option, i.e., if:

$$U(d, p_2, \theta) \geq -\mathbf{1}\{\theta \leq \theta_1\}\, p_2\, \beta\, L(d) \tag{2.8}$$

Consumers with $\theta < \theta_1$ who participate in the first period earn negative expected utility when they choose not to participate in period 2, for $\beta > 0$. Even if they do not participate, they will suffer harm $\beta\, L(d)$ with probability $p_2 \in \{p_n, p_b\}$. Thus, they find it **less** costly to use the service again, relative to the case of $\beta = 0$. Holding participation θ_1 fixed, this force will *raise* participation in *both* subgames n and b. To understand the impact on investment incentives for investment we need to ask how the difference $(\theta_n - \theta_b)$ is affected by changes in β. Importantly, a marginal change in β is going to a affect participation θ_b only when $\theta_b < \theta_1$ to begin with: that is precisely the case in which there is no "fresh demand" in subgame b and the decision of the marginal consumer of that subgame is affected by β. The same argument holds for θ_n. Thus, to sign the effect of β on $(\theta_n - \theta_b)$, we need to understand how thresholds $\theta_1, \theta_n, \theta_b$ are jointly determined in equilibrium.

If θ_1 is the indifferent consumer type in the first-period, and $\theta_1 \in (0,1)$, it must

[25] The EU GDPR (and its UK version) introduces a right for individuals to have personal data erased, see Article 17 of the UK GDPR.

satisfy the following condition:

$$U(d,p_1,\theta_1)+\delta\Big[p_1\max\{U(d,p_b,\theta_1),-p_b\beta\,L(d)\}+(1-p_1)\max\{U(d,p_n,\theta_1),-p_n\beta\,L(d)\}\Big]=$$
$$\delta\Big[p_1\max\{U(d,p_b,\theta_1),0\}+(1-p_1)\max\{U(d,p_n,\theta_1),0\}\Big] \qquad (2.9)$$

Only in this section, I am being explicit about $\delta > 0$, the discount factor, assumed common between firm and consumers.

The right-hand side of (2.9) is the expected payoff from off-path second-period decisions of θ_1. Parameter β does not appear in the right-hand side, because it does not affect an agent's payoff following no first-period participation.[26] Whether the terms in the right hand side are positive depends on how θ_1 compares to the corresponding second-period threshold, θ_n or θ_b. In the Appendix, I prove the following.

Lemma 5. *For any $\beta \in [0,1]$ and in any equilibrium with incomplete participation in the first period, i.e., with $\theta_1 < 1$, it must be that $U(d,p_b,\theta_1) < 0$. This implies that the participation cutoffs for subgames $\{1,b\}$ satisfy $\theta_1 \geq \theta_b$.*

According to the above, there can be no "fresh" demand in the second period following a breach, i.e., it can never be that $\theta_b > \theta_1$ in equilibrium. For $\beta = 0$, i.e., in the baseline model, it holds that $\theta_1 > \theta_b$: some first-period consumers do not return in subgame b. This means that the indifferent consumer θ_1 exits in subgame b, for $\beta = 0$, and, by continuity, this continues to hold for small values of $\beta > 0$. However, for any e_1, there exists sufficiently large β, such that all previously active consumers return in subgame b. In the extreme case of $\beta = 1$, all consumers with $\theta \leq \theta_1$ who have already shared their data with the firm find it "costless" to participate again. The above Lemma implies that in such equilibria,

[26] The baseline model corresponds to $\beta = 0$, in which case consumers' optimal decisions maximize each period's expected utility separately, and equation (2.9) yields the usual condition $U(d,p_1,\theta_1) = 0$.

it must be that $\theta_b = \theta_1$.

We can, thus, distinguish between two types of equilibria, which differ depending on whether $\theta_b = \theta_1$ or $\theta_b < \theta_1$. In any equilibrium, it must be that $\theta_n \geq \theta_b$. I show the following:

Proposition 5. *Starting at an equilibrium with $\theta_1 > \theta_b$, a marginal increase in β causes a reduction in first-period equilibrium investment, e_1 and a reduction in first-period participation θ_1.*

I provide the key intuition for this result and prove it formally in the Appendix. The impact of data retention on investment incentives is, determined by the direct effect of β on $\theta_n - \theta_b$, holding investment conjectures fixed. In turn, β affects participation only in second-period information sets in which there is *no* fresh demand. For example, if there is fresh demand in subgame n, i.e., $\theta_n > \theta_1$, the marginal consumer of that information set did not participate in period 1, hence is unaffected by changes in β. Thus, in that case, marginal changes in β do not have a direct effect on θ_n.[27]

At an equilibrium with $\theta_1 > \theta_b$, holding investment conjectures fixed, an increase in β means that participation in subgame b *rises*; the marginal consumer of that information set is also active in period 1 and for that reason affected by changes in β. As explained in the previous paragraph, participation in subgame n either increases or stays unchanged following changes in β, depending on whether or not there is fresh demand in that information set, i.e., on whether or not $\theta_n \geq \theta_1$. If $\theta_n > \theta_1$, the result follows, since β will not impact the indifferent θ_n user. If $\theta_n < \theta_1$, and there is a positive direct effect of β on both θ_n and θ_b, but the difference $(\theta_n - \theta_b)$ still decreases: the negative impact of β on returning consumers' outside option is of *greater* magnitude in $s = b$ than in $s = n$, because

[27] For the case of large sunk cost, i.e., $\theta_b = \theta_1$ equilibria, it can be shown that outside of edge cases, marginal changes in β leave both θ_n and θ_b *unaffected*, and thus do not affect investment incentives.

the probability of a breach in the former information set is larger. This causes a decrease in the difference $(\theta_n - \theta_b)$ and implies a reduction in the marginal benefit of investment.

To finish my analysis of this extension, I show how investment incentives differ between the baseline of $\beta = 0$ and the case of complete data retention, $\beta = 1$.

Proposition 6. *For sufficiently convex cost of investment: the equilibrium with the lowest investment under $\beta = 0$ involves larger investment than the equilibrium with the lowest investment under $\beta = 1$.*

This Proposition has immediate policy implications: endowing consumers with the right to withdraw their data from vendors they have previously shared it with will increase investment in data protection; this is precisely what the GDPR "right to erasure" imposes, and we should thus expect it to have increased firms' data protection investments. The empirical finding of Miller and Tucker [2018] is closely related: in their context of genetic testing data, endowing consumers with "property rights" over their medical information is found to *increase* take-up of the service. My model accordingly predicts an increase in first-period activity, not only because consumers are less reluctant to share data holding the firm's investment fixed but also because banning data retention increases equilibrium investment.

2.6 Limits on Data Collection

So far, I have considered a model with fixed data collection in every subgame and used it to assess the impact of existing privacy and cyber-security related policies. In this section, I ask a novel policy question that my model is uniquely suited to answer. I ask whether, and how, a regulator can raise consumer surplus by imposing *limits* on the amount of data that the firm can collect in the second

period of the model. In particular, I consider a regulator that can ex-ante *commit* to data limits at the beginning of period 1, and also condition these limits on whether or not the firm suffers a breach in the first period.

In general, the answer to these questions will depend on the initial level of data collection. To provide answers with clear policy relevance, I extend the baseline model in two different ways, corresponding to two different regimes of endogenous data collection, and ask the aforementioned questions at the equilibrium of each regime.

The two regimes differ in whether the firm or the consumers have control over data collection. In the regime of *consumer control*, active consumers can choose in every period (and information set) the amount of data they want to share with the firm. They can thus react to new information about the firm and adapt how much data they share with it to maximize their second-period expected utility. In the regime of *firm control*, the firm chooses its profit-maximizing data requirement in each subgame; I use control in the *ex-post* sense, meaning neither the firm (in the firm regime) nor the consumers (in the consumer regime) can pre-commit in the first period to their second-period decisions.

These regimes map accurately to how data collection is determined in reality: the EU GDPR enforced consumer "opt-out", giving consumers the choice of whether or not to share each specific piece of their personal data with a firm and the ability to opt out of each individual use of their data.[28] I capture this setting with the consumer regime.[29] In contrast, in the absence of such "opt-out" regulation, it is plausibly the case that firms choose how much data to collect in a profit maximizing manner.

[28]See for instance, GDPR.eu [2018], which, referring to Article 4, Recital 43 of the EU GDPR claims that consent must be item-specific and specific to each data-processing activity. Similar choice-granularity requirements are imposed by the UK ICO, see Information Commissioner's Office [2025b].

[29]In this regime, a planner that imposes limits on data "collection" essentially bars consumers from sharing as much data as they would find ex-post optimal to share.

2.6.1 Equilibrium in the two regimes

To facilitate analysis of this section, I introduce the following assumptions:

Assumption 2: $U(d, p, \theta)$ is strictly quasi-concave in d for every p, θ, with $U_{d,\theta} = 0$, and $U_{d,p} < 0$.

Assumption 3: $D(d, p) := G(\theta(d, p))$ is strictly quasi-concave in d for every d, p, and also has $D_{d,p} < 0$.

Assumption 4: Revenue per consumer $r(d)$ is log-concave.

Assumption 2 states that greater probability of a breach reduces the marginal value of data collection for consumers. Assumption 3 is an assumption on the cdf G, and is implied by Assumption 2 under uniform G. Assumption 4 bounds the rate at which per-consumer revenue increases in data and, alongside Assumption 3, guarantees that revenue $\Pi(d, p)$ is log-concave in data d (and thus quasi-concave).

In each regime, the timing of the game with endogenous data collection is almost identical to that of the baseline model. I will refer to $d \in [d^{min}, d^{max}]$ as the level of data *sharing* in the consumer-regime and as the level of data *collection* in the firm regime. In particular, I will denote by d_b data in subgame b and by d_n data in subgame n.

Consumer regime. In each information set, consumers choose both whether to be active users, and if so, the level of data to share with the firm, taking their current beliefs into account. Assumption $U_{d,\theta} = 0$ implies that consumers agree on the optimal value of data collection. For given probability of suffering a breach, active users share $d^C(p) := argmax_d\, U(d, p, \theta)$, which is uniquely defined by Assumption 2. The *negative* sign of $U_{d,p} < 0$ implies that $d^C(p)$ is *decreasing*. Finally, a lower p increases firm revenue via both greater demand and greater revenue per consumer. In other words, if we define $\Pi^C(p) := \Pi(d^C(p), p)$, we obtain the intuitive $d\Pi^C/dp < 0$, which implies a positive reputational premium

for the firm. An *equilibrium* under the consumer regime, $\{e_s^C, \mu_s^C, p_s^C, d_s^C, \theta_s^C\}$, for $s \in \{1, n, b\}$, identified using the superscript C, must satisfy the following:

1. $e_n^C = e_b^C = 0$ and e_1^C satisfies the investment f.o.c., given that revenue in each $s \in \{1, n, b\}$ is $\Pi^C(p_s^C)$.

2. Posterior beliefs are consistent with Bayes' rule, given investment level e_1^C. For $s \in \{n, b\}$, that means $\mu_s^C = \mu_s(e_1^C)$ and $p_s^C = p(\mu_s(e_1^C), 0)$.

3. Given e_1^C, active consumers choose data-sharing in each period and state according to $d_s^C = d^C(p_s^C)$, and $U(d_s^C, \theta_s^C, p_s^C) = 0$ where $s \in \{1, n, b\}$.

Similar to the baseline setting, in the case of multiplicity, I will refer to *equilibrium* as the pure-strategy (Perfect Bayesian) equilibrium that features the lowest value of first-period investment.

Firm regime. At the beginning of each period, the firm announces the level of data that a consumer must share with the firm in order to use the service. Importantly, the firm cannot influence consumer conjectures about investment in period one with the choice of d_1, because the expected marginal profit of e_1 does not depend on d_1. For that reason I will ignore d_1 for most of the following analysis. I define $d^F(p) := argmax_d\, \Pi(d, p)$ and $\Pi^F(p) := \Pi(d^F(p), p)$, and applying the Envelope Theorem yields $d\Pi^F/dp < 0$. Similar to above, I define equilibrium under firm-control, $\{e_s^F, \mu_s^F, p_s^F, d_s^F, \theta_s^F\}$, for $s \in \{1, n, b\}$.

The following result is derived by minimally varying the arguments for the proof of Proposition 1, and I omit the proof. I nevertheless state it for completeness.

Proposition 7. *There always exists a pure-strategy equilibrium in the unregulated consumer-regime and firm-regime games. For each regime, the pure-strategy equilibrium with the smallest first-period investment is stable. I will denote these investment levels by e_1^C and e_1^F. For sufficiently convex cost, equilibrium in each regime is unique.*

In the rest of this section, I consider in each regime a planner who plays first at the beginning of the game and ex-ante commits to **data limits** $\{d_n^{cap}, d_b^{cap}\}$ for the second period. In the firm regime, this means that the firm is restricted to choosing $d_n \leq d_n^{cap}$ in subgame n and $d_b \leq d_b^{cap}$ in subgame b, and in the consumer regime, consumers choices are similarly restricted by these second-period data caps. In the Appendix, I show that in each regime, when the caps are equal to the equilibrium values of d_b, d_n, there exists an equilibrium in which all endogenous quantities coincide with those of the unregulated equilibrium identified above.

I perform comparative statics with respect to each data cap around the equilibrium of each regime. For each regime, I seek to understand whether restricting data collection in either subgame can increase total consumer surplus.

Table 2.2: Notation: Limits to Data Collection

Symbol	**Definition**
$d^C(p)$	Consumer-optimal level of data sharing, given breach probability p
d_1^C, d_n^C, d_b^C	Data sharing in the unconstrained consumer-regime equilibrium
$d^F(p)$	Revenue-maximizing level of data collection, given breach probability p
d_1^F, d_n^F, d_b^F	Data sharing in the unconstrained firm-regime equilibrium
d^{cap}	Regulatory cap on data

2.6.2 Limits in the firm regime

In either regime, the imposition of data limits will induce a new equilibrium. The change in total consumer surplus will be due to three channels: (1) changes in equilibrium investment, holding data collected and consumer participation fixed, (2) changes in the amount of data collected, and (3) changes in consumer participation in the new equilibrium. Since participation decisions are always made (ex-post) optimally by consumers, the Envelope Theorem applies and there are

no first-order effects on consumer surplus via the third channel. I will refer to the first channel as the *indirect* effect of data restrictions and the second as the *direct* effect via reduced data collection.

$$\Delta CS \simeq \sum_{s\in\{1,n,b\}} \left[\frac{\partial CS_s}{\partial e_1}\Delta e_1 + \frac{\partial CS_s}{\partial d_s}\Delta d_s \right] \tag{2.10}$$

Starting from the unregulated equilibrium of the firm regime, consider a data cap that restricts only data collection following a breach, $d_b^{cap} < d_b^F$. In the new equilibrium, even though only the cap for d_b has changed, there may be different equilibrium values of d_n and d_1 too, if equilibrium investment changes. This is because e_1 affects consumers' posterior breach probabilities and thus the amount of data firm will optimally collect in each information set. I appeal to the following result to sign the direct effects of limiting data collection on consumer utility.

Lemma 6. *Fix the probability of a breach perceived by consumers, p, and assume that the revenue-maximizing data level is $d^F(p) < d^{max}$. If $U_{d,\theta} = 0$, it holds that $d^C(p) < d^F(p)$. Consumers' expected utility $U(d, p, \theta)$ is strictly decreasing in data at the level $d = d^F(p)$. The firm's revenue $\Pi(d, p)$ is strictly increasing in data at $d = d^C(p)$.*

According to this Lemma, if consumers agree on what their preferred level of data collection is, they would always rather that the firm asks for less data in each information set, holding investment fixed, and all direct effects in (2.10) are negative. The result is intuitive: the firm's decision cannot be profit maximizing if marginally increasing d would both increase demand and revenue per consumer. Next, I argue that restricting data collection in either subgame away from their initial values, d_n^F and d_b^F, will have no first-order impact on the firm's first-period investment.

Lemma 7. *(1) Fix the probabilities of breach in each second-period informa-*

tion set, p_n, p_b. A marginal increase in d_n increases equilibrium e_1 if and only if $d_n < d_n^F(p_n)$; a marginal increase in d_b increases equilibrium e_1 if and only if $d_b > d_b^F(p_b)$.
(2) Imposing small limits on either d_n or d_b relative to their firm-regime equilibrium values has no first-order effect on equilibrium investment, e_1.

The intuition behind the result is the following: at the initial equilibrium, the firm is collecting the (ex-post) optimal amount of data in each subgame of period two. Hence, small restrictions in the data collection of either subgame cause no first-order changes in the revenue of that subgame. Since data collection only affects investment incentives via the reputational premium $\Pi_n - \Pi_b$, there will be no first-order changes in first-period investment incentives. The fact that there is no first order impact on e_1 means that there is no first order impact on perceived posterior probabilities and thus on the equilibrium data collection levels d_1 and d_n. The only relevant direct effect is that via the reduction in d_b, and we saw in Lemma 6 that this effect is positive. Putting everything together, we obtain the following result:

Proposition 8. *Starting from the equilibrium of the firm regime with $d_b^{cap} = d_b^F$ and $d_n^{cap} = d_n^F$, a regulator can raise **total** consumer surplus by imposing small restrictions in the amount of data that firms can ask for in either information set of the second period, relative to the unregulated equilibrium.*

In the new equilibrium following either restriction, there are no first-order changes to consumers' first-period decisions and consumer surplus, and consumers are better off in period 2 because of the lower data collection.

2.6.3 Limits in the consumer regime

Let us now consider a regulator under the consumer regime. Given that there are no externalities between consumers, their participation and data-collection

decisions in the second period maximize consumer surplus of that period. Thus, holding investment fixed, there are no direct first-order effects of data limits on consumer surplus, and the direct effects in (2.10) are not of first order when evaluated at the unregulated equilibrium of the consumer regime. In contrast to the analysis of the firm regime, however, restricting consumers' data sharing *will* induce first-order changes in the firm's first-period investment. We can show the following:

Lemma 8. *Starting from the equilibrium of the consumer regime with* $d_b^{cap} = d_b^C$ *and* $d_n^{cap} = d_n^C$*, equilibrium investment* e_1 *increases if the planner commits to a data-sharing limit in subgame* b *that is marginally smaller than* d_b^C*, but it decreases if the planner commits to a limit* d_n^{cap} *marginally smaller than* d_n^C*.*

The intuition is similar to that of Lemma 7. Consider a tightening of the data cap d_b^{cap}. Since by Lemma 6, revenue is still increasing in d at the consumer-optimal level of data, a restriction in d_b decreases revenue in subgame b and thus increases the firm's incentive to invest in security and avoid a breach.

As already shown in Section 2.4, the impact of an increase in first-period investment e_1 on first-period consumer surplus is always positive, but according to Proposition 3, the marginal impact via learning on second-period consumer surplus qualitatively depends on the technology that is in place.[30] This has implications for the planner's problem at hand: if the impact of first-period investment e_1 on total consumer surplus is positive (negative), the planner will want to change data collection in a way that induces higher (lower) e_1.

Proposition 9. *Starting from the equilibrium of the consumer regime with* $d_b^{cap} = d_b^C$ *and* $d_n^{cap} = d_n^C$*:*

(a) if the marginal impact of first-period investment on total consumer surplus is

[30] The results of Section 2.4 were derived for the case of exogenously fixed data collection. In the proofs of the Propositions I present in this Section, I show how these results extend to the consumer regime.

positive, the planner can improve total consumer surplus by limiting second-period data collection following a breach relative to its equilibrium level, d_b^C. This will always be the case if investment enables learning.
(b) if investment impedes learning and the marginal impact of first-period investment on total consumer surplus is negative, the planner can improve total consumer surplus by limiting second-period data collection after no breach was observed relative to its equilibrium level, d_n^C.

In case (a), when investment enables learning, consumers would benefit if they could collectively commit to *greater* data sharing with high-reputation firms (i.e., firms who do not suffer a breach in period 1) and *lower* data sharing with low-reputation ones. But if investment impedes learning, there may be too much equilibrium investment from a consumer surplus perspective, in which case the appropriate policy would be to limit data sharing with high-reputation firms.

It is worth noting that because there are no data or participation externalities between consumers, if the planner could only announce data limits at the beginning of the second period rather than ex-ante commit to them in the first, he would not be able to achieve a consumer-surplus increase over the consumer-regime equilibrium.

Finally, before discussing the policy implications of the above analysis, it is useful to juxtapose the results in Lemmas 7 and 8 with the seemingly contradictory comparative statics result in Lemma 1 for the baseline model. The latter refers to the case in which we change both d_n and d_b starting from the same value $d_n = d_b = d$. However, this is not the comparative statics exercise performed in Lemmas 7 and 8. Here, I allow the planner to restrict data collection *differently* in each information set of the second period, by conditioning data-collection restrictions on the firm's first-period security outcome.

2.6.4 Policy implications and discussion

The results of this section have important policy implications. Proposition 8 reveals that some degree of opt-out rights is always beneficial to consumers, relative to the setting where firms make take-it-or-leave-it offers to consumers with respect to data collection. Consumers *should* be able to decline at least some kinds of data collection. On the other hand, Proposition 9 suggests appropriate data limits can also increase total consumer surplus relative to the consumer-regime equilibrium which is implemented by consent regulation. In case (a) of Proposition 9, he can do so by imposing restrictions, relative to equilibrium d_b^C, on data collection by firms that have recently suffered a data breach. In case (b) of Proposition 9, in which total consumer surplus will increase if first-period investment decreases, the policy prescription can be implemented by uniformly limiting data sharing with any firm to a level below d_n^C, which will not affect firms with low reputation in the second period, since $d_b^C < d_n^C$.

Before concluding this section, it is useful to briefly discuss Assumption 2 which states that expected utility net of the taste type θ, $(v(d) - p\,L(d))$, is quasi-concave in d for every p. The stricter assumption of concavity holds if $v''(d) < 0$ and $L''(d) > 0$.

Consumers' benefit $v(d)$ captures utility from better targeting with ads, or recommendations, or simply from being able to use more features of the service which require data collection. Decreasing marginal benefit to consumers is a natural assumption if the first data points deliver the most increases in targeting accuracy, or if consumers care about not being shown intrusive, irrelevant ads, which can be achieved even with little personalization.

Convex harm $L(d)$ can be justified in two ways: first, we could interpret higher levels of d as the firm collecting more *sensitive* data whose leakage is perceived as increasingly more harmful by privacy-concerned consumers. According to the

paper by de Matos and Adjerid [2022] finds that consumers consider some types of data to be more sensitive than others, for instance, geo-location data. Second, we could think of d as the quantity of similarly sensitive data: harm can be convex if malicious parties who get access to consumers' data via a data breach can harm consumers increasingly much as the amount of stolen data increases, by better tailoring phishing and identity theft schemes to their victims.

2.6.5 Heterogeneity in privacy preferences

The policy recommendation in the firm regime relies on the insight that the profit-maximizing firm collects too much data from an ex-post consumer surplus perspective. In turn, this insight is derived under the assumption of common consumer preferences over data-sharing, $U_{d,\theta} = 0$, but the recent empirical literature on data privacy, e.g. Aridor et al. [2023] and Lin [2022], finds heterogeneity in consumers' willingness to share data with firms. I want to understand the implications of such consumer heterogeneity for optimal policy prescriptions. I relax the assumption that $U_{d,\theta} = 0$ and instead assume that consumers with higher θ are more sensitive to potential data breaches, captured by $U_{d,\theta} < 0$ and $U_{p,\theta} < 0$. To be precise, I replace Assumption 2 with the following:

Assumption 5: (Privacy sensitivity) For any (d, p, θ), $U_\theta < 0$, $U_{\theta,d} < 0$, and $U_{\theta,p} < 0$.

Assumption 6: (Concavity) For any (d, p, θ), $U_{dd} < 0$.

For instance, $U(d, p, \theta) = v(d) - p\,\theta\,L(d)$ satisfies these assumptions. As this example shows, if consumers only differ in their sensitivity to data breaches, it follows that the most privacy sensitive active user is also the one with the lowest utility, i.e., the marginal user. Assumptions $U_\theta < 0$ and $U_{d,\theta} < 0$ guarantee that I only consider cases in which this is true.

I will focus on the implications of such heterogeneity for a regulator who faces

the **firm regime**. In each subgame the firm demands a **single** data level that each active consumer must provide. A consumer is active iff $U(d, p, \theta) \geq 0$ and these assumptions imply a unique cutoff type $\theta(d, p)$. When consumers have heterogeneous tastes with respect to data sharing, is it still the case that the firm asks for too much data in the second period, from an aggregate consumer surplus perspective? In other words, how is the conclusion of Lemma 6 affected?

Under our assumptions, there is a unique level of data that maximizes participation, $d^*(p)$, and a unique level that maximizes firm revenue, $d^F(p)$.[31] The following Lemma provides the key intuition behind the results of this section:

Lemma 9. *Given a probability of breach, $p \in (0, 1)$, assume that $d^{CS} < d^{max}$ is a value at which consumer surplus is maximized:*

- *$U_d(d^{CS}, \theta(d^{CS}, p), p) < 0$, the marginal consumer at d^{CS} prefers less data sharing at the margin. Participation is decreasing in data at any d^{CS} that maximizes consumer surplus.*
- *Conversely, consumer surplus is increasing at the unique d^* that maximizes consumer participation.*

To see why we obtain this result, assumption $U_{d,\theta} < 0$ implies that if U_d was weakly positive at point $(d, p, \theta(d, p))$ it would be strictly positive for all $\theta < \theta(d, p)$. Thus, increasing d would benefit the marginal consumer and all infra-marginal ones, too, and d cannot be CS-maximizing. So, the *marginal* consumers would rather that *less* data is collected at d^{CS}. This implies that a marginal increase in d from d^{CS} reduces participation. And at the level that maximizes *participation*, aggregate consumer surplus is *increasing* in data collection. This is true because the value d^* that maximizes participation, maximizes the data utility of the **marginal** user. But then, $U_{d,\theta} < 0$ means that all infra-marginal users still have positive marginal utility of data at d^*.

[31] I prove the first statement in the Appendix.

To relate the above discussion to the the value of data limits in the firm regime, note that by the same argument that led to Lemma 6, we know that $d^F > d^*$. However, under consumer heterogeneity, it is now ex-ante ambiguous whether consumer surplus is decreasing or not at d^F. If not, then imposing data caps will still improve welfare of almost-marginal types and *increase* participation but it will *reduce* aggregate consumer surplus.[32]

Whether consumer surplus is increasing at d^F will depend on the degree of consumer heterogeneity in privacy preferences, and on the firm's *business model*, as captured by function $r(d)$. In the benchmark of no heterogeneity in data preferences, the firm collects too much from a consumer surplus perspective. On the other hand, if $r'(d) = 0$, revenue-per-consumer is positive but independent of data sharing, the monopolist would optimally ask for the amount of data that maximizes *participation* and aggregate consumer surplus will not benefit from caps, according to the previous Lemma. This is due to a Spence [1975] quality distortion: when $r'(d) = 0$, $d = d^*$ maximizes the marginal consumer's expected utility and the firm's revenue. But the marginal consumer, $\theta(d, p)$, values privacy more than the infra-marginal ones, thus if the monopolist firm seeks solely to maximize participation, it over-provides *privacy* relative to the quantity that maximizes aggregate consumer surplus. By continuity this will continue to hold, when $r'(d) > 0$, for sufficiently small slope. I state the following result, for the case of linear revenue per consumer.

Lemma 10. *Fix some $p \in (0, 1)$, and assume $r(d) = r_0 + r_1 d$ with $r_0, r_1 \geq 0$,*

1. *For $r_0 > 0$ and r_1 sufficiently close to zero, consumer-surplus is increasing at the ex-post profit-maximizing data level.*

2. *If there is a unique maximizer of consumer surplus $d^{CS} < d^{max}$, and $d^F <$*

[32] The argument is the same as in the case without consumer heterogeneity: imposing small caps relative to equilibrium in $s = \{n, b\}$ raises aggregate CS if the direct utility effect on consumer surplus is positive, i.e., if $d^F(p_s^F) > d^{CS}(p_s^F)$, where p_s^F is the firm-regime equilibrium breach probability in $s = \{n, b\}$.

d^{CS} for any (r_0, r_1), that also holds for $r_0 = 0$.

3. *For $r_0 = 0$, if there is a unique maximizer of consumer surplus $d^{CS} < d^{max}$, $d^F > d^{CS}$ if and only if the (absolute) elasticity of participation with respect to data collection, $\varepsilon_{D,d}(d^{CS})$, is smaller than 1.*

The first point follows directly from the above discussion. The second point is intuitive: the monopolist firm is most likely to "over-collect" data relative to consumer surplus if it strongly values data volume instead of consumer participation. The policy prescription of Proposition 8 is overturned valid if data collection contributes sufficiently little to the firm's per-consumer revenue, relative to participation. I summarize the policy recommendations derived in this section in Table 2.3.

This extension has only focused on how heterogeneity affects optimal data caps in the firm regime. Preliminary analysis suggests that for the consumer regime the result of Proposition 9 generalizes in the following way: if consumers share different amounts of data, the maximum amount of data shared in the consumer regime is that shared by the *least* privacy-sensitive consumers. Capping data collection below that point will only affect these consumers, and raise e_1^*.

Table 2.3: Policy Recommendations

Regime	**Heterogeneity**	**Business Model**	**Policy**
Consumer	No	Any	Limit sharing with **breached** firms
Firm	No	Any	Limit collection for **either** history.
Firm	Yes	Data-heavy	Limit collection for **either** history.
Firm	Yes	Participation-heavy	Limits do **not** raise CS.

2.7 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. Uncertainty about the overall level of cyber risk is what motivates the firm to invest in security, but greater levels of investment may either impede or enable learning about the level of this risk, and in the former case negatively impact consumer surplus in the second period. The impact of investment on total consumer surplus can be non-monotonic, with implications for the use of penalties for firms that suffer data breaches. 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. In the consumer regime, data-sharing is chosen to maximize current-period consumer surplus. 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", if greater investment impedes learning about the true levels of cyber risk in the second period.

As shown in the extension to Section 6, heterogeneity of consumers with respect to privacy preferences may overturn the result derived for the firm regime. Extending the analysis of heterogeneous privacy preferences to the consumer regime is a natural next step.

Appendix 2.A Baseline model

2.A.1 Proof of Proposition 1

The solution concept is Perfect Bayesian Equilibrium. In equilibrium the following must be true:

1. $e_n^* = e_b^* = 0$, since zero effort is the firm's dominant strategy in each information set of the second period and consumers take that into account, along with posteriors $\mu_n(e_1^*)$ and $\mu_b(e_1^*)$ to optimally determine participation θ_n and θ_b.

2. Posterior reputations are formed using Bayes Rule and consumers make participation decisions in the first period optimally responding to e_1^*. The assumptions $f^h(1) > 0$ (high-risk type breached with positive probability even at maximal investment) and $f^\ell(0) < f^h(0) \leq 1$ guarantee that for $\mu \in (0,1)$ both possible outcomes occur on the equilibrium path for any e_1^* and posteriors are always well defined.

3. Given the expected participation of consumers in each information set of the second period, the firm chooses first-period investment to maximize total expected profit net of investment cost.

Fixing consumers' conjectures at $\tilde{e_1}$ implies posterior breach probabilities for the second period of $\tilde{p_n} = p(\mu_n(\tilde{e_1}), 0)$ and $\tilde{p_b} = p(\mu_b(\tilde{e_1}), 0)$. The firm's first-period investment is unobserved by consumers and thus cannot influence the conjecture

$\tilde{e_1}$, which the firm must take as fixed in its profit maximization problem. For given conjecture, the firm's marginal expected revenue from increasing first-period investment is found by partially differentiating (2.3):

$$MB(e_1, \tilde{e_1}) := \overbrace{-\frac{\partial p(\mu_1, e_1)}{\partial e_1}}^{(+)} \underbrace{\Big(\Pi(d, \tilde{p_n}) - \Pi(d, \tilde{p_b})\Big)}_{\text{Reputational Premium}} \tag{2.11}$$

Differentiating with respect to e shows that if $f^r(e)$ is weakly convex for each r, i.e., increases in investment have a lower impact on reducing the probability of a breach, then the firm's second-order condition is satisfied at a stationary point (sufficient condition). In an interior equilibrium, $e_1^* \in (0, 1)$, it must be that $\tilde{e_1} = e_1^*$ and the above quantity must equal the marginal cost of investment $C'(e_1^*)$. I will refer to the *equilibrium* marginal benefit of e_1 as the quantity:

$$MB^{eq}(e_1) := MB(e_1, e_1) = -\frac{\partial p(\mu_1, e_1)}{\partial e_1}\Big(\Pi\big(d, p(\mu_n(e_1), 0\big) - \Pi\big(d, p(\mu_b(e_1), 0\big)\Big) \tag{2.12}$$

This is the marginal benefit of investment e_1 when consumers anticipate this level of investment and the reputational premium is adjusted accordingly. To proceed, I distinguish two cases, depending on whether equilibrium marginal benefit crosses marginal cost of first-period investment.

Case A: A crossing point exists.

Each intersection of $MB^{eq}(e_1)$ and $C'(e_1)$ is an equilibrium investment level. Assumption 1 guarantees that when consumers anticipate first-period investment of $e_1 = 0$, $MB^{eq}(0) > 0$. This is because Assumption 1 implies that at $e_1 = 0$, $\mu_n(0, \mu) > \mu > \mu_b(0, \mu)$, i.e., absence of a breach is evidence of the low-risk type. Under the assumption of $C'(0) = 0$, this means that $e_1 = 0$ cannot be an equilibrium. At the smallest e_1 that satisfies $MB^{eq}(e_1^*) = C'(e_1^*)$, it must be true that $(MB^{eq})'(e_1^*) < C''(e_1^*)$. Thus, the smallest equilibrium e_1^* is stable. Sufficiently

steep $C'(e_1)$ would also ensure uniqueness.

Note that an intersection can only occur at e_1^* such that $f^\ell(e_1^*) < f^h(e_1^*)$. If that were not the case, then $\mu_b(e_1^*) > \mu_n(e_1^*)$ and *breach* would be evidence of low risk. Hence, $p_n > p_b$ and the reputational premium would be negative.

Case B: No crossing point exists.
As already argued, Assumption 1 and $C'(0) = 0$ imply that this can only be the case if $MB^{eq}(e_1) > C'(e_1)$ for all $e_1 \in [0, 1]$. The unique equilibrium in this case is $e_1^* = 1$. This case is ruled out by sufficiently steep marginal cost of investment. It is also ruled out whenever the technology is such that $f^\ell(e_1) > f^h(e_1)$ for any value $e_1 \in (0, 1]$. At such a value, $\mu_n(e_1) < \mu_b(e_1)$, thus $MB^{eq}(e_1) < 0$ and continuity implies that a point $MB^{eq}(e) = C'(e)$ would then exist. Thus, such an equilibrium is possible only when $f^\ell(e) < f^h(e)$ for all e.

2.A.2 Proof of Lemma 1

We have shown in Proposition 1 that either the unique equilibrium is at $e_1^* = 1$, or an interior equilibrium exists and in particular, a stable equilibrium e_1^* exists at the lowest crossing point of marginal benefit and marginal cost curves. For the interior equilibrium case, consider the comparative statics of e_1^*. I appeal to the Implicit Function Theorem, to implicitly differentiate equation $MB^{eq}(e_1) - C'(e_1) = 0$ that defines an interior equilibrium, with $MB^{eq}(e_1)$ defined in 2.12. At the stable equilibrium we have identified in Proposition 1, an increase in d will cause an increase in e_1^* iff: To show the above, it suffices to show that the cross-partial derivative of revenue is negative for all $p \in (p(\mu_n(e_1^*), 0), p(\mu_b(e_1^*), 0))$.

$$\frac{\partial \Pi}{\partial p} = r(d)\frac{\partial \theta(d, p)}{\partial p} < 0 \tag{2.13}$$

$$\frac{\partial^2 \Pi(p, d)}{\partial d \partial p} = r'(d)\frac{\partial \theta(d, p)}{\partial p} + r(d)\frac{\partial^2 \theta(d, p)}{\partial d \partial p} \tag{2.14}$$

Notice that $\frac{\partial^2\theta(d,p)}{\partial d\partial p} = \frac{\partial^2 U(d,p,\theta)}{\partial d\partial p} = -L'(d) < 0$ and that guarantees the desired sign for the right-hand side if $\theta \sim U$.

2.A.3 Proof of Lemma 2

We have shown in Proposition 1 that either the unique equilibrium is at $e_1^* = 1$, or an interior equilibrium exists and in particular, a stable equilibrium e_1^* exists at the lowest crossing point of marginal benefit and marginal cost curves. For the second case, which is always the case for not too low marginal cost, consider the comparative statics of the interior equilibrium e_1^*. I prove the result, operating under the stated assumption $f^\ell(e) > 0$ for all e, which guarantees that neither outcome is ever perfectly revealing of the firm's type (and thus both posteriors depend on the prior μ).

Define:

$$-\frac{\partial p(\mu, e)}{\partial e} = -\mu\frac{df^\ell(e)}{de} - (1-\mu)\frac{df^h(e)}{de} =: s(e,\mu) > 0 \tag{2.15}$$

and

$$\frac{\partial s(e,\mu)}{\partial \mu} = -\frac{\partial^2 p(\mu, e)}{\partial \mu e} = \frac{df^h(e)}{de} - \frac{df^\ell(e)}{de} =: \phi(e) \tag{2.16}$$

which may be positive or negative. Any interior equilibrium must satisfy the first-order condition:

$$H(e_1^*, \mu) := s(e_1^*, \mu)\Delta\Pi(\mu, e_1^*) - C'(e_1^*) = 0 \tag{2.17}$$

where $\Delta\Pi(\mu, e_1^*) := \Pi(p_n(\mu_n, 0)) - \Pi(p_n(\mu_b, 0))$ and I am being explicit about the dependence of second-period profits on the prior μ and consumers' first-period investment conjectures via μ_n, μ_b. I have defined e_1^* as the smallest root of this first-order condition and argued in Proposition 1 that it corresponds to a stable equilibrium, meaning $H_e(e_1^*, \mu) := \partial H/\partial e < 0$ at $e = e_1^*$. By continuity of the

marginal benefit and cost functions, there exists a continuous function $e_1^*(\mu)$, with slope:

$$\frac{\partial e_1^*}{\partial \mu} = \frac{-1}{H_e(e_1^*, \mu)}\Big[s(e_1^*, \mu)\frac{\partial \Delta\Pi(\mu, e_1^*)}{\partial \mu} + \phi(e_1^*)\Delta\Pi(\mu, e_1^*)\Big] \tag{2.18}$$

To show that e_1^* is a quasi-concave function of the ex-ante probability of a low-risk type, μ, I will show that the slope is decreasing in μ at every stationary point of the function $e_1^*(\mu)$.

$$\frac{\partial^2 e_1^*}{\partial \mu^2} = \frac{-1}{H_e^2}\Bigg[\frac{\partial}{\partial \mu}\left(s(e_1^*, \mu)\frac{\partial \Delta\Pi}{\partial \mu} + \phi(e_1^*)\Delta\Pi(e_1^*, \mu)\right)H_e - \underbrace{\left(s(e_1^*, \mu)\frac{\partial \Delta\Pi(\mu, e_1^*)}{\partial \mu} + \phi(e_1^*)\Delta\Pi(\mu, e_1^*)\right)}_{=0 \text{ by } (2.18)} H_{e,\mu}\Bigg]$$

but at any stationary point, the subtracted quantity inside the brackets must be zero. I have also ignored changes in the numerator via e_1^*, since that is (up to the first order) constant at the points we are looking at. So, we are left with:

$$\begin{aligned}\frac{\partial^2 e_1^*}{\partial \mu^2} &= \frac{-1}{H_e}\left[\frac{\partial}{\partial \mu}\left(s(e_1^*, \mu)\frac{\partial \Delta\Pi}{\partial \mu} + \phi(e_1^*)\Delta\Pi(e_1^*, \mu)\right)\right] \\ &= \frac{-1}{H_e}\left(s(e_1^*, \mu)\frac{\partial^2 \Delta\Pi}{\partial \mu^2} + 2\phi(e_1^*)\frac{\partial \Delta\Pi}{\partial \mu}\right)\end{aligned}$$

and I claim that the above quantity is negative, proving the quasi-concavity statement. To show this, I will first show that each of the two terms in parentheses are negative. Starting with the second term, at any set of parameters that satisfies $\partial e_1^*/\partial \mu = 0$, it must be by (2.18) that: so that:

$$2\phi(e_1^*)\frac{\partial \Delta\Pi}{\partial \mu} = -2\frac{(\phi(e_1^*))^2\Delta\Pi(\mu, e_1^*)}{s(e_1^*, \mu)} < 0$$

because $s(e,\mu) > 0$ for all e, μ and $\Delta\Pi(e_1^*, \mu) > 0$ for any equilibrium e_1^*. Now, for the first term in brackets, the direct effect of the prior belief on second-period profits is via posterior beliefs' impact on participation:

$$\begin{aligned}\frac{\partial \Pi(p_n)}{\partial \mu} &= r(d)\frac{\partial G(\theta_n)}{\partial \theta_n}\frac{\partial \theta_n}{\partial p_n}\frac{\partial p_n}{\partial \mu} \\ &= -r(d)L(d)g(\theta_n)\frac{\partial p_n}{\partial \mu} > 0\end{aligned}$$

where $\partial p_n/\partial \mu < 0$ because low-types are breached with lower probability in period 2 and posterior μ_n increases in μ. Differentiating again:

$$\frac{\partial^2 \Pi(p_n)}{\partial \mu^2} = -r(d)L(d)\Big[g(\theta_n)\frac{\partial^2 p_n}{\partial \mu^2} + g'(\theta_n)\frac{\partial \theta_n}{\partial p_n}(\frac{\partial p_n}{\partial \mu})^2\Big]$$

where the second inequality follows because the indifferent consumer type is linear in the belief μ_n. Similarly, we obtain:

$$\frac{\partial^2 \Pi(p_b)}{\partial \mu^2} = -r(d)L(d)\Big[g(\theta_b)\frac{\partial^2 p_b}{\partial \mu^2} + g'(\theta_b)\frac{\partial \theta_b}{\partial p_b}(\frac{\partial p_b}{\partial \mu})^2\Big]$$

and subtracting:

$$\begin{aligned}\frac{\partial^2 \Delta\Pi}{\partial \mu^2} &= \frac{\partial^2 \Pi(p_n)}{\partial \mu^2} - \frac{\partial^2 \Pi(p_b)}{\partial \mu^2} \\ &= -r(d)L(d)\Big[g(\theta_n)\underbrace{\frac{\partial^2 p_n}{\partial \mu^2}}_{(+)} - g(\theta_b)\underbrace{\frac{\partial^2 p_b}{\partial \mu^2}}_{(-)} + g'(\theta_n)\overbrace{\frac{\partial \theta_n}{\partial p_n}}^{(+)}(\frac{\partial p_n}{\partial \mu})^2 - g'(\theta_b)\overbrace{\frac{\partial \theta_b}{\partial p_b}}^{(+)}(\frac{\partial p_b}{\partial \mu})^2\Big]\end{aligned}$$

As already argued in the main text, in any equilibrium, it must hold that $f(e^*, \ell) < f(e^*, h)$. In that case, the above signs of the posterior beliefs' second derivatives obtain.

Lemma 11. *It always holds that the first-order partial derivatives of μ_n, μ_n with respect to μ are positive. For any e such that $f^\ell(e) < f^h(e)$, it also holds that*

$\frac{\partial^2 \mu_n}{\partial \mu^2} < 0$ *and* $\frac{\partial^2 \mu_b}{\partial \mu^2} > 0$.

This immediately shows that the negative sign obtains for uniformly distributed consumer types, when $g'(\theta) = 0$, and will continue to hold whenever the slopes are not too large.

The technical assumption $f(1, h) > 0$ implies that as $\mu \to 1$, the two posteriors remain well-defined and $\mu_n, \mu_b \to 1$. Thus, the equilibrium marginal benefit curve becomes horizontal at zero. In other words, the reputation premium is completely eroded and any breaches are interpreted as originating from an unlucky low-risk type. As $\mu \to 0$ but while μ is still positive, the equilibrium marginal benefit approaches zero for any value of e_1, because $\mu_n, \mu_b \to 0$. in which case the lack of breach will be interpreted as the result of investment by a high-risk type. Equilibrium investment e_1^* arbitrarily close to zero for extreme values of the prior μ and $e_1^* > 0$ for intermediate μ imply that at least one stationary point of $\partial e_1^*(\mu)/\partial \mu$ exists and the second-order condition we establishes implies that this point is unique.

Appendix 2.B Consumer Surplus

2.B.1 Proof of Proposition 2

I repeat equation (2.5) from the main text:

$$CS_1(e_1, d) := \int_0^{\theta(d, p_1(e_1))} U\big(d, p_1(e_1), \theta\big)\, dG(\theta) \tag{2.19}$$

Taking the first derivative of CS_1 yields:

$$\frac{\partial CS_1(e_1, d)}{\partial e_1} = \underbrace{\frac{\partial CS}{\partial p_1}}_{<0} \underbrace{\frac{\partial p_1}{\partial e_1}}_{<0} + \underbrace{\frac{\partial CS}{\partial \hat{\theta}}}_{=0} \frac{\partial \theta(d, p)}{\partial e_1} > 0 \tag{2.20}$$

The partial derivative with respect to participation is zero because the indifferent consumer has expected utility equal to zero. This is equal to:

$$\frac{\partial CS_1(e_1,d)}{\partial e_1} = G(\hat{\theta})\frac{\partial \hat{\theta}}{\partial p_1}\underbrace{\frac{\partial p_1}{\partial e_1}}_{<0} = -G(\hat{\theta})L(d)\frac{\partial p_1}{\partial e_1} \tag{2.21}$$

and differentiating again yields:

$$\frac{\partial^2 CS_1(e_1,d)}{\partial (e_1)^2} = g(\hat{\theta})L^2(d)\left(\frac{\partial p_1}{\partial e_1}\right)^2 - G(\hat{\theta})L(d)\frac{\partial^2 p_1}{\partial (e_1)^2} \tag{2.22}$$

which is true when f and therefore p are linear in e (or when f and p are concave).

2.B.2 Proof of Proposition 3

I will prove the results for slope and convexity for the case when investment impedes learning. For the case in which investment enables learning, the steps of the proof are very similar. Throughout this proof, I will refer to conditions (2.28) and (2.33), such that investment impedes learning, which I provide and discuss in the later proof of Proposition 4. The reader can first refer to the proof of that result but I will make explicit when steps referenced here have been proven in that section.

For the following proofs I will omit the d argument from functions. I remind the reader the definition of CS_2 from the main text:

$$CS_2(e_1) := E_{\mu,e_1}\left[\int_0^{\theta(d,p_s(e_1))} U\big(d,p_s(e_1),\theta\big)\,dG(\theta)\right] \tag{2.23}$$

For the following proof, it is also helpful to define the function:

$$CS(\hat{\theta},r) := \int_0^{\hat{\theta}} U(d,f^r(0),\theta)dG(\theta) \tag{2.24}$$

This is the consumer surplus achieved in period 2, when participation is $\hat{\theta}$ and the firm's type is r. Note that the probability of a breach is $f^r(0)$ because there neither type invests in that period.

Lemma 12. *If* (2.28) *and* (2.33) *hold for all* $e_1 \in (0,1)$*, ex-ante second-period consumer surplus,* $CS_2(e_1)$ *is a decreasing function of investment.*

Proof.

$$\begin{aligned} CS_2(e_1) =& \mu\big[f^\ell(e_1)CS(\theta_b,\ell) + (1-f^\ell(e_1))CS(\theta_n,\ell)\big] \\ &+ (1-\mu)\big[f^h(e_1)CS(\theta_b,h) + (1-f^h(e_1))CS(\theta_n,h)\big] \end{aligned}$$

where $\theta_n = \theta(p_n(e_1),d)$ and $\theta_b = \theta(p_b(e_1),d)$. When differentiating with respect to e_1, the effects via consumers' optimal decisions θ_b, θ_n are not of first-order, and what remains is:

$$\begin{aligned} \frac{dCS_2(e_1)}{de_1} =& -\mu\frac{df^\ell(e_1)}{de}\big[CS(\theta_n,\ell) - CS(\theta_b,\ell)\big] \\ &- (1-\mu)\frac{df^h(e_1)}{de}\big[CS(\theta_n,h) - CS(\theta_b,h)\big] \end{aligned} \tag{2.25}$$

The difference in the first brackets is positive and the second is negative: to see this, note that the first difference is the aggregate expected utility of users with $\theta \in (\theta_b,\theta_n]$ when they are active and they face an ℓ type in period 2: all of these users derive positive expected utility when the posterior reputation of the firm is $\mu_n < 1$ and the posterior probability of a breach is $p_n = p(\mu_n(e_1),0)$, so they must also obtain (strictly) positive expected utility if they face an ℓ type with certainty, i.e., the probability of breach is $f^\ell(0) < p_n$. Using an identical argument we see that the second pair of brackets contains a negative difference: users with θ_b obtain zero expected utility when posterior beliefs are μ_b, so they must obtain strictly negative expected utility when $r = h$ with certainty, and so will users with even

higher θ. I will show that the total right-hand side is *negative* under (2.28) and (2.33).

$$\frac{dCS_2(e_1)}{de_1} = -\mu \frac{df^{\ell}(e_1)}{de} \int_{\theta_b}^{\theta_n} \underbrace{\frac{\partial CS(\theta, \ell)}{\partial \theta}}_{(+)} d\theta - (1-\mu) \frac{df^{h}(e_1)}{de} \int_{\theta_b}^{\theta_n} \underbrace{\frac{\partial CS(\theta, h)}{\partial \theta}}_{(-)} d\theta$$

where the signs are implied by the same arguments as above, and putting everything under a single integral:

$$\frac{dCS_2(e_1)}{de_1} = -\int_{\theta_b}^{\theta_n} \left[\mu \frac{df^{\ell}(e_1)}{de} \underbrace{\frac{\partial CS(\theta, \ell)}{\partial \theta}}_{(+)} + (1-\mu) \frac{df^{h}(e_1)}{de} \underbrace{\frac{\partial CS(\theta, h)}{\partial \theta}}_{(-)} \right] d\theta$$

If the integrand is positive *for all* $\theta \in (\theta_b, \theta_n)$, the result is true. Towards a contradiction, assume that for at least for some $\theta \in (\theta_b, \theta_n)$, the integrand is **negative**. This means that for some $\theta \in (\theta_b, \theta_n)$, the following is negative:

$$\begin{aligned} & \mu \frac{df^{\ell}(e_1)}{de} \frac{\partial CS(\theta, \ell)}{\partial \theta} + (1-\mu) \frac{df^{h}(e_1)}{de} \frac{\partial CS(\theta, h)}{\partial \theta} \\ = & \underbrace{\frac{df^{h}(e_1)}{de}}_{(-)} \left[\mu \left(\frac{\frac{df^{\ell}(e_1)}{de}}{\frac{df^{h}(e_1)}{de}} \right) \frac{\partial CS(\theta, \ell)}{\partial \theta} + (1-\mu) \frac{\partial CS(\theta, h)}{\partial \theta} \right] \end{aligned}$$

which in turn implies that the bracketed expression is **positive**. For any $\theta \in$

(θ_b, θ_n), we know the signs of the consumer surplus partial derivatives.

$$
\begin{aligned}
0 < \mu\left(\frac{\frac{df^{\ell}(e_1)}{de}}{\frac{df^{h}(e_1)}{de}}\right)\overbrace{\frac{\partial CS(\theta,\ell)}{\partial\theta}}^{(+)} + (1-\mu)\overbrace{\frac{\partial CS(\theta,h)}{\partial\theta}}^{(-)} < \quad \text{(by (2.33))} \\
\mu\left(\frac{f^{\ell}(e_1)}{f^{h}(e_1)}\right)\frac{\partial CS(\theta,\ell)}{\partial\theta} + (1-\mu)\frac{\partial CS(\theta,h)}{\partial\theta} = \\
\frac{1}{f^{h}(e_1)}\left[\mu f^{\ell}(e_1)\frac{\partial CS(\theta,\ell)}{\partial\theta} + (1-\mu)f^{h}(e_1)\frac{\partial CS(\theta,h)}{\partial\theta}\right] = \\
\frac{p(\mu,e_1)}{f^{h}(e_1)}\left[\frac{\mu f^{\ell}(e_1)}{p(\mu,e_1)}\frac{\partial CS(\theta,\ell)}{\partial\theta} + \frac{(1-\mu)f^{h}(e_1)}{p(\mu,e_1)}\frac{\partial CS(\theta,h)}{\partial\theta}\right] = \quad \text{(by (2.27))} \\
\frac{p(\mu,e_1)}{f^{h}(e)}\left[\mu_b(e_1)\frac{\partial CS(\theta,\ell)}{\partial\theta} + (1-\mu_b(e_1))\frac{\partial CS(\theta,h)}{\partial\theta}\right] = \\
\frac{p(\mu,e_1)}{f^{h}(e_1)}\Big[\mu_b(e_1)U(\theta,\ell) + (1-\mu_b(e_1))U(\theta,h) - \theta\Big]g(\theta) < 0
\end{aligned}
$$

which yields a contradiction, so that the integrand is negative for no value of $\theta \in (\theta_b, \theta_n)$. The contradiction stems from the fact that the last bracketed expression is, by definition, set to zero by $\theta_b = \theta(p_b(e_1), d)$ and is negative for any $\theta > \theta_b$ (for consistency with CS in this proof, I am also skipping the d argument of U). □

Lemma 13. *Assume $f_{ee} = 0$, i.e., the technology is linear in investment. Then, under (2.28) and (2.33), ex-ante second-period consumer surplus, $CS_2(e_1)$ is a* ***convex*** *function of investment.*

Proof. To obtain the convexity of $CS_2(e_1)$, differentiate (2.25) again (with linear technology $f_{ee} = 0$):

$$
\begin{aligned}
\frac{d^2CS_2(e_1)}{(de_1)^2} = & -\mu\frac{df^{\ell}(e_1)}{de}\left[\overbrace{\frac{\partial CS(\theta_n,\ell)}{\partial\theta}}^{(+)}\frac{\partial\theta_n}{\partial e} - \overbrace{\frac{\partial CS(\theta_b,\ell)}{\partial\theta}}^{(+)}\frac{\partial\theta_b}{\partial e}\right] \\
& -(1-\mu)\frac{df^{h}(e_1)}{de}\left[\underbrace{\frac{\partial CS(\theta_n,h)}{\partial\theta}}_{(-)}\frac{\partial\theta_n}{\partial e} - \underbrace{\frac{\partial CS(\theta_b,h)}{\partial\theta}}_{(-)}\frac{\partial\theta_b}{\partial e}\right] \qquad (2.26)
\end{aligned}
$$

The two conditions we operate under imply that $\partial\theta_n/\partial e < 0$ and $\partial\theta_b/\partial e > 0$. Thus, by arguments made in the previous paragraphs of this proof, the first expression in big brackets is negative: the positive effect of e_1 on CS_2 of more frequent good outcomes by the ℓ type becomes smaller. The quantity in the second pair of brackets is positive and the negative effect of more frequent good outcomes by the h type also becomes of smaller magnitude. The right-hand side of (2.26) seems to be of ambiguous sign, but we can re-arrange into:

$$\frac{d^2CS_2(e_1)}{(de_1)^2} = -\overbrace{\frac{\partial\theta_n}{\partial e_1}}^{(-)}\left[\mu\frac{df^\ell(e_1)}{de}\frac{\partial CS(\theta_n,\ell)}{\partial\theta} + (1-\mu)\frac{df^h(e_1)}{de}\frac{\partial CS(\theta_n,h)}{\partial\theta}\right] + \underbrace{\frac{\partial\theta_b}{\partial e_1}}_{(+)}\left[\mu\frac{df^\ell(e_1)}{de}\frac{\partial CS(\theta_b,\ell)}{\partial\theta} + (1-\mu)\frac{df^h(e_1)}{de}\frac{\partial CS(\theta_b,h)}{\partial\theta}\right]$$

The proof by contradiction at the last step of the proof for Lemma 12 applies again, and both expressions in big brackets are shown to be *strictly negative*, hence the overall second derivative is positive. □

2.B.3 Proof of Proposition 4

The formulas for the two posterior beliefs are:

$$\mu_b(e_1) = \frac{\mu f^\ell(e_1)}{\mu f^\ell(e_1) + (1-\mu)f^h(e_1)}, \quad \mu_n(e_1) = \frac{\mu(1-f^\ell(e_1))}{\mu(1-f^\ell(e_1)) + (1-\mu)(1-f^h(e_1))} \tag{2.27}$$

For linear technology, Assumption 1 implies $h_0 > \ell_0$ and $h_0 - h_1 e \geq \ell_0 - \ell_1 e$, for $e > 0$. The remaining restriction is that $1 \geq f(e, r_0, r_1) \geq 0$ for both $r \in \{h, \ell\}$.

2.B.3.1 Investment impedes learning

The posterior following "no breach" is decreasing in e_1 if:

$$\frac{\partial \mu_n}{\partial e} < 0 \iff \tag{2.28}$$

$$\frac{\partial}{\partial e}\Big[\frac{1-f^h(e)}{1-f^\ell(e)}\Big] > 0 \iff \tag{2.29}$$

$$-\frac{df^h(e)}{de}(1-f^\ell(e)) - (-\frac{df^\ell(e)}{de})(1-f^h(e)) > 0 \iff \tag{2.30}$$

$$\frac{df^\ell(e)}{de}(1-f^h(e)) > \frac{df^h(e)}{de}(1-f^\ell(e)) \iff \tag{2.31}$$

$$\frac{\frac{df^\ell(e)}{de}}{\frac{df^h(e)}{de}} < \frac{1-f^\ell(e)}{1-f^h(e)} \tag{2.32}$$

and the last inequality follows because $\frac{df^h(e)}{de} < 0$. The second condition included in Definition 2 is that:

$$\frac{\partial \mu_b}{\partial e} > 0 \iff \tag{2.33}$$

$$\frac{\partial}{\partial e}\frac{f^h(e)}{f^\ell(e)} < 0 \iff \tag{2.34}$$

$$\frac{df^h(e)}{de}f^\ell(e) - \frac{df^\ell(e)}{de}f^h(e) < 0 \iff \tag{2.35}$$

$$\frac{df^\ell(e)}{de}f^h(e) > \frac{df^h(e)}{de}f^\ell(e) \iff \tag{2.36}$$

$$\frac{\frac{df^\ell(e)}{de}}{\frac{df^h(e)}{de}} < \frac{f^\ell(e)}{f^h(e)} \tag{2.37}$$

Under the linearity assumption, conditions (2.28) and (2.33) that are part of the "impeding" definition do not depend on the value of e. Under Assumption 1, condition (2.33) implies condition (2.28) and, rewriting (2.33), we obtain the necessary and sufficient condition for investment to impede learning:

$$\frac{\ell_1}{h_1} < \frac{\ell_0}{h_0} \tag{2.38}$$

By the second part of Assumption 1, the right-hand side is smaller than 1, so that $h_1 > \ell_1$ is necessary but not sufficient.

2.B.3.2 Investment enables learning

Reversing the inequalities from the previous section, the conditions that must hold for an increase in e to spread posteriors out are:

$$\frac{\partial \mu_n}{\partial e} > 0 \iff \frac{\frac{df^\ell(e)}{de}}{\frac{df^h(e)}{de}} > \underbrace{\frac{1-f^\ell(e)}{1-f^h(e)}}_{\geq 1} \tag{2.39}$$

$$\frac{\partial \mu_b}{\partial e} < 0 \iff \frac{\frac{df^\ell(e)}{de}}{\frac{df^h(e)}{de}} > \underbrace{\frac{f^\ell(e)}{f^h(e)}}_{\leq 1} \tag{2.40}$$

The inequalities in underbraces are again implied by Assumption 1. Conditions 2.39 and 2.40 do not depend on the value of e and the former implies the latter. The necessary and sufficient condition is:

$$\frac{\ell_1}{h_1} > \frac{1-\ell_0}{1-h_0} \tag{2.41}$$

and the right-hand side is larger than 1 so that $\ell_1 > h_1$ is necessary but not sufficient.

Appendix 2.C Minimum Security Standards

2.C.1 Proof of Proposition 4

When minimum security standards $\underline{e}$ are in place, the posterior breach probability in subgame $s \in \{n, b\}$ is $p(\mu_s(e_1), \underline{e})$. For that reason, the equilibrium marginal benefit of investment depends on both e_1 and $\underline{e}$ and I will refer to the

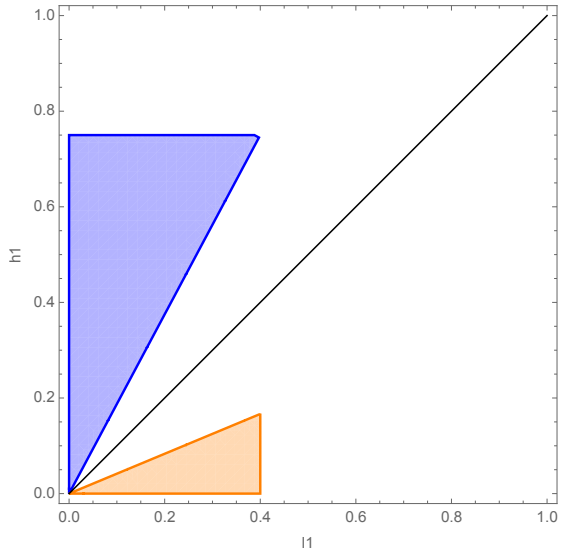


Figure 2.5: For linear technologies $f^h(e) = 0.75 - h_1 e$, $f^\ell(e) = 0.4 - \ell_1 e$. Horizontal axis is ℓ_1 and vertical axis is h_1. The blue region corresponds to parameters such that (2.28) and (2.33) hold for all $e \in [0,1]$ (impedes learning) and the orange region represents parameters such that (2.39) and (2.40) hold for all $e \in [0,1]$ (enables learning). Given that $h_0 > \ell_0$ the necessary condition for investment to impede learning is $h_1 > \ell_1$, and learning can only be enabled if $\ell_1 > h_1$.

marginal benefit curve $MB^{eq}(e_1; \underline{e})$. There are two types of equilibria, depending on whether or not the constraint is binding. In an unconstrained equilibrium, $e_1^* > \underline{e}$ and the first-order condition that must be satisfied in an equilibrium is $MB(e_1; \underline{e}) = MC(e_1)$. In a constrained equilibrium, the relevant condition that must be satisfied is $MB^{eq}(\underline{e}; \underline{e}) \leq MC(\underline{e})$.

Proof of (a)**:** Investment impedes learning.

At an unconstrained equilibrium with $e_1 < \underline{e}$, e_1 is defined by the first-order condition (with equality) and is locally differentiable in $\underline{e}$. When investment impedes learning, $MB^{eq}(e_1; \underline{e})$ is strictly decreasing in e_1 and thus there is a unique, stable equilibrium. To obtain the sign of the derivative of e_1 with respect to $\underline{e}$, we differentiate the marginal benefit with respect to $\underline{e}$. $MB^{eq}(e_1, \underline{e}) = r(d)k_1\Big(\Pi(d, p(\mu_n(e_1), \underline{e})) - \Pi(d, p(\mu_b(e_1), \underline{e}))\Big)$ where $k_1 := \mu\ell_1 + (1-\mu)h_1 > 0$. Omitting the positive constant $\frac{\partial p_1}{\partial e_1} = k_1 > 0$ (under the linear technology), the

impact of $\underline{e}$ on marginal expected profit is:

$$\begin{aligned}\frac{\partial}{\partial \underline{e}}\big(\Pi(p_n) - \Pi(p_b)\big) &= r(d)\Big(g(\theta_n)(-1)L(d)\frac{\partial p_n}{\partial \underline{e}} - g(\theta_b)(-1)L(d)\frac{\partial p_b}{\partial \underline{e}}\Big) \\ &= r(d)L(d)\Big(g(\theta_n)A - g(\theta_b)B\Big)\end{aligned}$$

where:

$$A = \mu_n \ell_1 + (1-\mu_n)h_1 > 0$$

$$B = \mu_b \ell_1 + (1-\mu_b)h_1 > 0$$

and $A - B = (\mu_n - \mu_b)(\ell_1 - h_1)$ which has the sign of $(\ell_1 - h_1)$ since in any interior equilibrium $(\mu_n - \mu_b) > 0$.

$$\frac{\partial}{\partial \underline{e}}\big(\Pi(p_n) - \Pi(p_b)\big) < 0 \iff \frac{A}{B} < \frac{g(\theta_b)}{g(\theta_n)} \tag{2.42}$$

We know from Proposition 4, that a necessary condition for investment to impede learning is that $h_1 > \ell_1$, in which case $A < B$ and the sign is negative under the uniform distribution and a sufficient condition is that $g(\theta_b) > g(\theta_n)$ or that g is weakly decreasing. This proves the first part of the statement.

For the second part of the statement in (a), start from an equilibrium in which the constraint binds. A necessary condition is that $MB^{eq}(\underline{e};\underline{e}) \leq MC(\underline{e})$. I show that introducing $\underline{e}' > \underline{e}$ means that the new equilibrium satisfies $e_1 = \underline{e}'$. Because investment impedes learning, p_n is increasing in e_1 and p_b is decreasing in e_1. Hence $MB^{eq}(e_1;\underline{e})$ is decreasing in e_1. We also just showed that under our assumptions, MB^{eq} is decreasing in $\underline{e}$. Thus, $MB^{eq}(\underline{e}';\underline{e}') < MB^{eq}(\underline{e};\underline{e}) \leq C'(\underline{e}) < C'(\underline{e}')$ and the firm has no incentive to raise first-period investment further from the minimum requirement of $\underline{e}'$.

Proof of (b)**:** Investment enables learning.

We focus on sufficiently convex cost to guarantee existence of a unique and stable equilibrium. Following arguments similar to the first part (a), we see that:

$$\frac{\partial}{\partial \underline{e}}\big(\Pi(p_n) - \Pi(p_b)\big) > 0 \iff \frac{A}{B} > \frac{g(\theta_b)}{g(\theta_n)} \tag{2.43}$$

When investment enables learning, it must be that $h_1 < \ell_1$, in which case $A > B$ and a sufficient condition for the inequality to hold is that $g(\theta_b) \leq g(\theta_n)$, i.e., that g is weakly increasing. For the second part of (b): unlike in (a), an increase in $\underline{e}$ starting from a constrained equilibrium might mean that the new equilibrium e_1 overshoots the new $\underline{e}$. The reason is that when investment enables learning (which implies $\ell_1 > h_1$), the quantity $MB^{eq}(e_1; \underline{e})$ is increasing in both e_1 and $\underline{e}$.

Appendix 2.D Ban on Data Retention

2.D.1 Preliminary Lemma

Before proving the results stated in the main text, I argue that for any fixed e_1, all three participation thresholds are uniquely defined. I start with the second-period thresholds θ_n and θ_b. Consider subgame b. A user with $\theta > \theta_1$ participates if and only if $U(d, p_b, \theta) \geq 0$. A user with $\theta \leq \theta_1$ participates if and only if $U(d, p_b, \theta) \geq -\beta\, p_b\, L(d)$. By $U_\theta < 0$, if θ_1 does not participate, neither does any user with higher type. There are three possibilities:

1. The threshold is $\theta_b > \theta_1$ and type θ_b is defined by $U(d, p_b, \theta) = 0$. This case is ruled out by Lemma 5.

2. The threshold type θ_b is infra-marginal in $s = 1$, i.e., $\theta_b < \theta_1$. This means that there is some exit in subgame $s = b$. Threshold θ_b is uniquely defined by $U(d, p_b, \theta) = -\beta\, p_b\, L(d)$.

3. θ_b is exactly equal to θ_1 and $0 > U(d, p_b, \theta_1) > -\beta\, p_b\, L(d)$. In this case, there is no indifferent type in subgame b. Type $\theta_b + \epsilon$ has a larger outside option than the threshold type θ_b, by virtue of not having participated in $s = 1$.

The same argument and three possibilities arise for participation θ_n and both are uniquely defined for given e_1 and θ_1. The threshold θ_1 is determined from the indifference condition stated in the main text, which I repeat here:

$$U(d, p_1, \theta_1) + \delta\Big[p_1 \max\{U(d, p_b, \theta_1), -p_b\beta\, L(d)\} + (1 - p_1)\max\{U(d, p_n, \theta_1), -p_n\beta\, L(d)\}\Big] = \delta\Big[p_1 \max\{U(d, p_b, \theta_1), 0\} + (1 - p_1)\max\{U(d, p_n, \theta_1), 0\}\Big] \quad (2.44)$$

This equation does not depend on θ_n, θ_b. Both sides are decreasing in θ_1 and it is easy to verify that the slope of the left-hand side is smaller than -1 whereas the slope of the right-hand side is larger than $-\delta > -1$. Hence, the solution θ_1 (and we assume one exists) is uniquely defined. Additionally, higher β reduces the option value of not using the service in period two, thus decreasing the value of the left-hand side, and leading to the following result:

Lemma 14. *For any given $\beta \in [0, 1]$: fix any level of e_1, and the implied probabilities p_1, p_n, p_b. There exists a unique first-period indifferent type θ_1, and it is weakly decreasing in β. There exist unique participation thresholds $\theta_n \geq \theta_b$.*

2.D.2 Proof of Lemma 5

I show that there can be no equilibrium in which $\theta_1 < \theta_b$, i.e., no equilibrium with fresh demand in subgame b. Notice that $p_1 < p_b$, the posterior breach probability in subgame b is larger than the prior, for any equilibrium e_1. This means that in the absence of data retention, $\theta_1 > \theta_b$. Under data retention $\beta > 0$, I show that a customer that participates in $s = b$ must also have wanted to participate in $s = 1$.

Towards a contradiction, assume $\theta_1 < \theta_b$. Then the fresh consumer θ_b is indifferent in $s = b$ and by $U_\theta < 0$, we obtain:

$$0 = U(d, p_b, \theta_b) < U(d, p_b, \theta_1) \tag{2.45}$$

the last step also implies, by $U_p < 0$ and $p_n < p_b$, that:

$$U(d, p_n, \theta_1) > U(d, p_b, \theta_1) > 0 \tag{2.46}$$

Thus, type θ_1 obtains positive utility from participating in either subgame b or n and indifference equation (2.9) reveals that:

$$U(d, p_1, \theta_1) = 0 \tag{2.47}$$

Intuitively, the sunk cost, which reduces the outside option of early participating types, does not influence the decisions of type θ_1 if that type obtains positive utility in both future subgames. But for type $\theta_b > \theta_1$, equation (2.45) implies:

$$0 = U(d, p_b, \theta_b) < U(d, p_1, \theta_b) < U(d, p_1, \theta_1) = 0 \tag{2.48}$$

and the two inequalities are implies by $U_p < 0$, $U_\theta < 0$. We have obtained a contradiction. Hence, there is no equilibrium with $\theta_b > \theta_1$.

2.D.3 Proof of Proposition 5

I focus on the case in which equilibrium features $\theta_1 > \theta_b$.

Assuming stability of the initial equilibrium, standard use of the implicit function theorem implies that the sign of $\partial e_1^*/\partial\beta$ is determined by the marginal impact of β on the difference $(\theta_n - \theta_b)$, holding investment fixed.

In this case the first-period marginal user *exits* in subgame b, and earns their

negative expected outside utility, because they have already participated and part of their data is left behind. This means:

$$U(d, p_b, \theta_b) = -\beta\, p_b\, L(d) < 0 \tag{2.49}$$

Since $\theta_1 > \theta_b$, type θ_b participates in $s = 1$ and faces sunk cost in subgame b; thus, his type is defined uniquely by $U(d, p_b, \theta_b) = -\beta\, p_b\, L(d)$.

$$\frac{\partial \theta_b}{\partial \beta} = p_b L(d) > 0 \tag{2.50}$$

To find θ_n, we must make further progress. $U(d, p_b, \theta_b) = -\beta\, p_b\, L(d) < 0 \implies U(d, p_b, \theta_1) < 0$ by $U_\theta < 0$ and $\theta_1 > \theta_b$. The indifference condition (2.9) becomes:

$$\begin{aligned} U(d, p_1, \theta_1) + \delta\Big[- p_1 \beta p_b L(d) + (1 - p_1) \max\{U(d, p_n, \theta_1), -p_n \beta L(d)\}\Big] = \\ \delta\Big[(1 - p_1) \max\{U(d, p_n, \theta_1), 0\}\Big] \end{aligned} \tag{2.51}$$

I take cases according to the relation between θ_1 and θ_n.

- **Case 1:** $\theta_n > \theta_1$: the marginal $s = n$ consumer type does not participate in period 1, hence θ_n satisfies $U(d, p_n, \theta_n) = 0$. Threshold θ_n is locally unaffected by changes in β. Hence, the difference $(\theta_n - \theta_b)$ clearly decreases following a marginal increase in β.

- **Case 2:** $\theta_n < \theta_1$: the marginal $s = n$ consumer participates in period 1, and is defined by $U(d, p_n, \theta_n) = -\beta p_n L(d) < 0$.

  $$\frac{\partial \theta_n}{\partial \beta} = p_n\, L(d) > 0 \text{ and } \frac{\partial(\theta_n - \theta_b)}{\partial \beta} = (p_n - p_b) L(d) < 0 \tag{2.52}$$

  As a comment, note that in such an equilibrium, type θ_1 does not participate in $s = n$ and participates in $s = 1$ to obtain the first-period (positive) expected utility in spite of guaranteed negative expected utility in period 2.

- **Case 3:** $\theta_n = \theta_1$: the threshold consumer types for $s = 1$ and $s = n$ coincide. In such an equilibrium, it must be that θ_1 earns weakly negative utility from participation in subgame n, i.e., $0 > U(d, p_n, \theta_1)$. Otherwise, type $\theta_1 + \epsilon$, would also participate in $s = n$, for sufficiently small ϵ. Given $0 \geq U(d, p_n, \theta_1) > U(d, p_b, \theta_1)$, condition (2.51) yields:

$$U(d, p_1, \theta_1) + \delta\Big[- p_1 \beta p_b L(d) + (1 - p_1) U(d, p_n, \theta_1)\Big] = 0$$

and substituting terms and differentiating yields:

$$\frac{\partial \theta_n}{\partial \beta} = \frac{\partial \theta_1}{\partial \beta} = - \frac{p_1 p_b \delta L(d)}{1 + \delta(1 - p_1)} < 0 \tag{2.53}$$

$\partial \theta_b / \partial \beta > 0$, meaning that $(\theta_n - \theta_b)$ again decreases with marginal increases in β.

Thus, in any stable equilibrium that features $\theta_b < \theta_1$, marginal increase in β induces a marginal decrease in e_1^*.

2.D.4 Proof of Proposition 6

To prove this result, I will show that at every conjecture of consumers e_1, the expected marginal benefit of investment is greater under $\beta = 1$ than under $\beta = 0$.

Fix some e_1 level of conjectures and thus the associated probabilities p_1, p_b, p_n. The desired comparison boils down to whether the difference $(\Pi_n - \Pi_b)$ is greater for $\beta = 0$ or $\beta = 1$.

For $\beta = 0$, the threshold $\theta_b^{\beta=0}$ is given by the root to $U(d, p_b, \theta) = 0$ and the threshold $\theta_n^{\beta=0}$ is given by the root to $U(d, p_n, \theta) = 0$.

For $\beta = 1$, any equilibrium must fall into the case of $\theta_b = \theta_1$, see Lemma 5. There are two cases to consider, either $\theta_n > \theta_b = \theta_1$, or $\theta_n = \theta_b = \theta_1$.[33] It is easy to

[33]The second case corresponds to a setting in which the first-period indifferent type earns

see that the proposition holds for an equilibrium of the second type: there is no reputational premium if $\theta_n = \theta_b$ and thus the only candidate equilibrium value is $e_1^* = 0$. Since by Proposition 1 investment is always positive for $\beta = 0$, the result obtains.

For the case of $\theta_n > \theta_b = \theta_1$, the threshold type in subgame n is "fresh" demand and thus $\theta_n^{\beta=1}$ is also given by the root of $U(d, p_n, \theta) = 0$, hence $\Pi_n^{\beta=0} = \Pi_n^{\beta=1}$.

Additionally, we have derived in the proof of the previous result in this Section that (focusing on interior θ_1) that $U(d, p_1, \theta_1^{\beta=1}) > 0 > U(d, p_b, \theta_1^{\beta=1})$ and $\theta_b^{\beta=1} = \theta_1^{\beta=1}$ lies in between the roots of $U(d, p_b, \theta) = 0$ and $U(d, p_1, \theta) = 0$. Because $p_1 < p_b$ (the perceived probability of a breach is lower in $s = 1$ than in $s = b$ following a breach, for any e_1) and thus $U(d, p_b, \theta) < U(d, p_1, \theta)$, and $U_\theta < 0$, the previous statement implies that $\theta_b^{\beta=1}$ is *greater* than $\theta_b^{\beta=0}$ and the firm's revenue in subgame b is *greater* under $\beta = 1$ than under $\beta = 0$. This is intuitive: any type that participates in subgame b under $\beta = 0$ does so with outside option of zero. These types will also participate in subgame b under $\beta = 1$ when their outside option is lower (by $\theta_b \leq \theta_1$), but under $\beta = 1$ there is some *additional* returning types due to the sunk cost.

Combined with the statement about Π_n, we have shown that the marginal benefit of investment is strictly higher under $\beta = 0$, for any e_1. Thus, the first crossing point of $MB^{\beta=1}$ with MC is at a smaller e than the first crossing point of $MB^{\beta=0}$, proving the result in Proposition.

negative utility in both subgames of period 2.

Appendix 2.E Limits on Data Collection

2.E.1 Proof of Lemma 6

For the first part of the Lemma, note for a given breach probability p, the first-order condition for profit-maximizing data collection is:

$$\frac{\partial \Pi(d,p)}{\partial d} = \underbrace{r'(d)G(\theta(d,p))}_{\substack{\text{more data from} \\ \text{inframarginal users}}} + r(d)\underbrace{g(\theta(d,p))\frac{\partial \theta(d,p)}{\partial d}}_{\substack{\text{change in} \\ \text{participation}}} = 0$$

Assumptions 3 and 4 guarantee that profit is quasi-concave hence the maximizer must satisfy this first-order condition. For given p, denote by $d^F(p)$ the unique maximizer of $\Pi(d,p)$, and by $d^C(p)$ the maximizer of $U(d,p,\theta)$, which is unique by Assumption 2 (U is quasi-concave in d) and independent of θ, because $U_{d,\theta} = 0$. Since $r'(d) > 0$, the firm cannot be profit-maximizing unless the rightmost derivative is negative. Otherwise, an increase in d would surely raise revenue, yielding a contradiction. Thus, revenue is maximized at a point where participation is *decreasing* in d. Because $U_{d,\theta} = 0$, this implies all consumers' marginal utility in data is negative at that level. Hence $d^F(p) > d^C(p)$. If at some p, $d^C(p) = d^{max}$, it must be that $\frac{\partial u(d,p)}{\partial d} > 0$ for all d, hence $d^F(p) = d^{max}$, too.

2.E.2 Intermediate result: Equilibrium equivalence

In both the firm- and consumer-regime, the regulator plays first and ex-ante commits to caps d_b^{cap}, d_n^{cap} of data that can be collected/shared in period two. Once the regulator has set the data limits, the timing of the game in each regime is identical to that of the game without a regulator.

Firm regime In the equilibrium of the firm regime, $d_n = d_n^F$ and $d_b = d_b^F$. I show that when the data caps used by the planner in the firm regime are $d_n^{cap} = d_n^F$

and $d_b^{cap} = d_b^F$, there exists an equilibrium in which all endogenous quantities (data, investment, participation) induced coincide with those of the firm regime. Fix consumer conjectures to e_1^F and thus posterior breach probabilities perceived by consumers to p_n^F, p_b^F. The firm can do no better in information set b than to choose $d_b^* = d^F(p_b^F) = d_b^F$ and similarly, $d_n^* = d_n^F$. Thus, $\Pi_b^* = \Pi_b^F$ and $\Pi_n^* = \Pi_n^F$.

As in the baseline model, when the firm chooses e_1 it treats consumers conjectures and thus the posterior breach probabilities p_b, p_n as fixed: hence, the firm's FOC is $-p_1'(e_1)\Big(\Pi_n^* - \Pi_b^*\Big) - C'(e_1) = 0$. Because the investment-FOC in the unconstrained firm-regime equilibrium is satisfied by e_1^F when data collection is d_b^F and d_n^F, this FOC is also satisfied at investment level e_1^F when the regulators sets data caps at those same values. The SOC is always satisfied because investment does not directly affect the expected marginal benefit of the firm but marginal cost is increasing. Hence, we have identified an equilibrium of the game under these specific data caps.

Consumer regime Similarly, I show that when the regulator chooses limits to data sharing $d_n^{cap} = d_n^C$ and $d_b^{cap} = d_b^C$, there exists an equilibrium of the consumer regime with data-sharing limits in which all endogenous quantities (data, investment, participation) induced coincide with those of the firm regime.

Fix consumer conjectures to e_1^C and thus posterior breach probabilities perceived by consumers to $p_n^C = p_n(e_1^C)$, $p_b^C = p_b(e_1^C)$. Because d_b^C is ex-post optimal for consumers, i.e., $d_b^C = d^C(p_b)$, consumers can do no better in information set b than to choose $d_b^* = d_b^C$, and the associated participation level, and similarly in information set n. The firm's investment FOC must hold again at e_1^C, because it holds at the consumer-regime equilibrium without the constraints, and the SOC holds too. Hence, this is an equilibrium of the consumer regime with limits $d_n^{cap} = d_n^C$ and $d_b^{cap} = d_b^C$.

2.E.3 Proof of Lemma 7

In the firm regime with data caps d_n^{cap}, d_b^{cap}, quasi-concavity of revenue with respect to d implies that for any fixed breach probabilities p_b, p_n, the firm will choose $d_b^*(p_b) = \min\{d_b^{cap}, d^F(p_b)\}$ and $d_n^*(p_n) = \min\{d_n^{cap}, d^F(p_n)\}$. This implies values for the ex-post optimal revenue in each information set of period 2, $\Pi(p_b, d_b^{cap})$ and $\Pi(p_n, d_n^{cap})$.

As usual, the FOC that defines an equilibrium interior value of investment is:

$$MB^{eq}(e_1^*;\, d_n^{cap},\, d_b^{cap}) - C'(e_1^*) = 0 \tag{2.54}$$

where $MB^{eq}(e_1;\, d_n^{cap},\, d_b^{cap})$ is the firm's *equilibrium* marginal benefit of investment, for given data limits d_n^{cap}, d_b^{cap} and when posterior probabilities are correctly determined using e_1.

$$MB^{eq}(e_1;\, d_n^{cap},\, d_b^{cap}) = -\frac{\partial p_1}{\partial e_1}\left(\Pi(p_n(e_1), d_n^{cap}) - \Pi(p_b(e_1), d_b^{cap})\right) \tag{2.55}$$

where I emphasize that e_1 determines p_n, p_b. To obtain the desired result, I will show that even though d_n^*, d_b^* are not differentiable at investment levels where the data constraints switch between being slack or binding, the function MB^{eq} has indeed everywhere continuous partial derivatives in both e_1 and d_b^{cap}.

Step 1 First, I show partial differentiability with respect to d_b^{cap} (same arguments hold for d_n^{cap}). For given p_b, there exists an ex-post optimal value $d^F(p_b)$ and as already mentioned, $d_b^* = \min\{d_b^{cap}, d^F(p_b)\}$. For $d_b^{cap} < d_b^F \implies d_b^* = d_b^{cap}$, the partial derivative of $\Pi(p_b, d_b^{cap})$ exists and is given by:

$$\frac{\partial \Pi(p_b, d_b^{cap})}{\partial d_b^{cap}} = \frac{\partial \Pi}{\partial d_b^*}\overbrace{\frac{\partial d_b^*}{\partial d_b^{cap}}}^{=1}$$

and the limit as $d_b^{cap} \to d_b^F$ is 0, by the Envelope Theorem. For $d_b^{cap} > d_b^F \implies d_b^* = d_b^F$, the partial derivative of $\Pi(p_b, d_b^{cap})$ exists and is equal to zero, since d_b^* is locally unaffected by the data limit. Hence, the limit of the right and left-hand side partial derivatives of $\Pi(p_b, d_b^{cap})$ as $d_b^{cap} \to d_b^F$ coincide and the partial derivative at $d_b^{cap} = d_b^F$ exists. At points where the data cap binds exactly, $d_b^F = d_b^{cap}$, the derivative is zero.

Step 2 Following a similar argument, I show differentiability with respect to e_1. Take for example, some $\hat{e}$ such that $\hat{p_b} = p_b(\hat{e})$ and such that $d_b^* = d_b^{cap}$ in $[\hat{e}, \hat{e} + \delta)$ but $d_b^* = d^F(\hat{p_b})$ in $[\hat{e} - \delta, \hat{e})$, for some $\delta > 0$. $\Pi(p_b, d_b^{cap})$ is continuously differentiable with respect to e_1 both intervals. Our goal is to show that the derivative also exists at $e_1 = \hat{e}$. On either side of $\hat{e}$, the derivative is given by:

$$\frac{d\Pi(p_b(e_1), d_b^{cap})}{de_1} = \left[\frac{\partial \Pi}{\partial p_b} + \frac{\partial \Pi}{\partial d_b^*}\frac{\partial d_b^*}{\partial p_b}\right]\frac{\partial p_b}{\partial e_1} \tag{2.56}$$

Changes in the investment perception of consumers affects p_b which has both a direct effect on revenue, via changes in consumers' participation holding data collection fixed, and an indirect, via changing the firm's optimal data collection, d_b^*. In $[\hat{e}, \hat{e} + \delta)$, $d_b^* = d_b^{cap}$ and $\partial d_b / \partial e_1 = 0$. In $[\hat{e} - \delta, \hat{e})$, $d_b^* = d^F(\hat{p_b})$ and by the Envelope Theorem, $\partial \Pi_b / \partial d_b = 0$ at the ex-post optimum $d_b^* = d^F(\hat{p_b})$. Hence, in both intervals:

$$\frac{d\Pi(p_b(e_1), d_b^{cap})}{de_1} = \frac{\partial \Pi}{\partial p_b}\frac{\partial p_b}{\partial e_1} \tag{2.57}$$

and the continuity of $d_b^*(e_1)$ implies that this function is continuous at $\hat{e}$. The same will apply at every point e_1 such that either constraint switches from being binding to not and vice-versa. Hence, MB^{eq} is continuously differentiable.

Step 3 Given this differentiability, we can implicitly differentiate the equilibrium first-order condition to see how a change in the data-cap away from d_b^F will affect equilibrium investment, around the initial equilibrium induced by $\{d_n^{cap} =$

$d_n^F,\, d_b^{cap} = d_b^F\}$:

$$\frac{\partial e_1}{\partial d_b^{cap}} = -\frac{\partial MB^{eq}(e_1^F;\, d_n^F,\, d_b^F)/\partial d_b^{cap}}{\partial\big(MB^{eq}(e_1^F;\, d_n^F,\, d_b^F) - C'(e_1^F)\big)/\partial e_1} = 0 \tag{2.58}$$

since $\partial\Pi(p_b, d_b^{cap})/\partial d_b^{cap} = 0$ at $d_b^{cap} = d_b^F$. Intuitively, small changes in the data collection around the ex-post optimum yield no first-order changes in Π_b and thus no first-order changes in first-period investment incentives.

2.E.4 Proof of Proposition 8

Argument follows immediately from the previous Lemma and from the fact that there is a positive, first-order, direct effect on consumer utility from a reduction in d_b, given Lemma 6.

2.E.5 Proof of Lemma 8

Consider the consumer regime with limits to data sharing in period two. Given e_1, and thus $p_n(e_1), p_b(e_1)$, the equilibrium condition for e_1, given a choice of $\{d_b^{cap}, d_n^{cap}\}$ by the planner, is defined as $H(e_1; d_b^{cap}, d_n^{cap}) := MB^{eq}(e_1; d_b^{cap}, d_n^{cap}) - C'(e_1) = 0$. If the data caps are initially set at unconstrained consumer equilibrium values, $d_b^{cap} = d_b^C$ and $d_n^{cap} = d_n^C$, we know by the argument in the Intermediate Result of this Appendix that $H(e_1^C; d_b^C, d_n^C) = 0$. Quasi-concavity of expected utility in d implies that the amount of data shared by constrained consumers in subgame b is $d_b^*(e_1) = \min\{d^C(p_b(e_1)), d_b^{cap}\}$. The key difficulty in proving the present Lemma is that at this initial equilibrium, $d_b^*(e_1)$ is non-differentiable in the data cap. So at points where the constraint binds exactly, d_b^* is not differentiable in d_b^{cap} and the partial derivative of $\Pi(p_b(e_1), d_b^{cap})$ with respect to the data cap does not exist either.[34]

[34] And neither do the derivatives with respect to p_b or e_1, whose effects go through $d^C(p_b)$ and thus through the non-differentiable d_b^*.

Step 1 I assume that the initial equilibrium of this game is stable, i.e., that the equilibrium marginal benefit curve of investment crosses marginal cost from above at the equilibrium value e_1^C. This is guaranteed by the firm's cost of investment being sufficiently convex. In turn, stability of the initial equilibrium implies that the function H is locally decreasing in e_1. Therefore, for any sufficiently small $\delta > 0$, $H(e_1^C + \delta;\, d_b^C, d_n^C) < 0$ and $H(e_1^C - \delta;\, d_b^C, d_n^C) > 0$.

Step 2. I show that a decrease in the cap reduces H at e_1^C. I contemplate a decrease in the data limit, hence use the left partial derivative of d_b^*, $\partial d_b^*/d_b^{cap} = 1$. Then, I use the well-defined partial **left**-derivative of H with respect to the data limit d_b^{cap}, where $\partial d_b^*/\partial d_b^{cap} = 1$:

$$\frac{\partial H(e_1^C, d_b^C; d_n^C)}{\partial d_b^{cap}} = \frac{\partial MB^{eq}}{\partial d_b^{cap}} = \underbrace{p_1'(e_1^C)}_{(-)} \underbrace{\frac{\partial \Pi(p_b(e_1^C), d_b^C)}{\partial d_b^C}}_{(+)} < 0 \tag{2.59}$$

The first sign is negative because investment reduces the probability of a breach. We know the second sign because an increase in data collection from d_b^C raises the firm's revenue in subgame b, by Lemma 6. This implies:

$$H(e_1^C; d_b^C - \epsilon, d_n^C) > 0 \tag{2.60}$$

Step 3 Show the new equilibrium e_1 satisfies $e_1^{new} > e_1^C$. We already know that $H(e_1^C + \delta, d_b^C) < 0$. For small enough ϵ, i.e., for small enough restriction in data collection, it must also be that $H(e_1^C + \delta, d_b^C - \epsilon) < 0$, by continuity of H. Hence, continuity of H and the Intermediate Value Theorem imply that there exists e_1^{new} larger than e_1^C that satisfies the equilibrium FOC at the new data cap.

2.E.6 Proof of Proposition 9

In equations (2.5) and (2.6), I have defined consumer surplus of periods 1 and 2 as a function of data collected and investment e_1 for the case of fixed data collection, d, in every subgame. Extending the definitions in the obvious way, we can approximate the first-order change in total CS induced by a change in d_b^{cap} via the total differential around equilibrium:

$$\Delta CS \simeq \sum_{s\in\{1,n,b\}} \left[\frac{\partial CS_s}{\partial e_1}\Delta e_1 + \frac{\partial CS_s}{\partial \theta_s}\Delta\theta_s + \frac{\partial CS_s}{\partial d_s}\Delta d_s \right] \tag{2.61}$$

where I have decomposed the change in expected consumer surplus in each of the three information sets $s \in \{1, n, b\}$, into three components: (1) direct the impact via changing investment e_1, holding participation and data sharing decisions fixed, (2) the impact via changing participation decisions only, and (3) the impact via changing data sharing decisions only. Around the initial equilibrium where data and participation are chosen ex-post optimally by consumers, only effect (1) is of first order. From Lemma 8 we know that a restriction of data sharing in subgame b induces $\Delta e_1 > 0$, but a restriction of data sharing in subgame n will induce $\Delta e_1 < 0$. Hence all that is left is to sign $\frac{\partial CS}{\partial e_1}$. I show that the following holds:

Lemma 15. *Around an equilibrium of the consumer regime, the direct effect of first-period investment on second-period expected consumer surplus is negative, if investment impedes learning, and positive if investment enables learning.*

The above lemma follows from adapting the arguments that lead to the result in Proposition 3, and is proven in the Additional Material section of this Appendix.

Appendix 2.F Heterogeneity in Privacy Preferences

I remind the reader of the assumptions I make in this section:

Assumption 5: (Privacy sensitivity) For any (d, p, θ), $U_\theta < 0$, $U_{\theta,d} < 0$, and $U_{\theta,p} < 0$.

Assumption 6: (Concavity) For any (d, p, θ), $U_{dd} < 0$. Under these assumptions, I first establish the uniqueness of d^*, the data level that maximizes participation.

Lemma 16. *Fix some breach probability p. Under Assumptions 4 and 5, there exists a unique $d^*(p)$ that maximizes the mass of participating consumers.*

Proof. The indifferent consumer is defined by $U(d, p, \theta(d, p)) = 0$. I will show that $\theta(d, p)$ is a quasi-concave function of d, at any p. Implicit differentiation of the definition yields:

$$\frac{\partial \theta(d,p)}{\partial d} = -\frac{U_d}{U_\theta} \tag{2.62}$$

thus, participation threshold $\theta(d, p)$ is maximized at an interior d, it must be that:

$$U_d(d^*, p, \theta(d^*, p)) = 0 \tag{2.63}$$

Differentiating the condition $\theta_d U_\theta + U_d = 0$ with respect to d yields:

$$\begin{aligned} &\theta_{dd} U_\theta + \theta_d U_{d\theta} + U_{\theta\theta}(\theta_d)^2 + U_{dd} + U_{d\theta}\theta_d = 0 \iff \\ &\theta_{dd} = -\frac{U_{dd}}{U_\theta} < 0 \end{aligned}$$

where I have evaluated the above at any d that satisfies the first order condition and thus $\theta_d = 0$. We have shown that at any point that satisfies the first-order condition, the second-order condition holds too. Hence $\theta(d, p)$ is quasi-concave in d. □

2.F.1 Proof of Lemma 9

For any fixed (d, p), aggregate consumer surplus is maximized by extending participation up to the indifferent consumer, $\theta(d, p)$. Under the assumption that

participation is always positive and incomplete, i.e., that the indifferent consumer type satisfies $0 < \theta(d,p) < 1$, $\theta(d,p)$ is differentiable in both arguments and defined by $U(d,p,\theta(d,p)) = 0$. Given participation of consumers with $\theta \in [0, \theta(d,p)]$, consumer surplus in a given information set is a function of (d,p):

$$CS(d,p) := \int_0^{\theta(d,p)} U(d,p,\theta) g(\theta) d\theta \tag{2.64}$$

The first-order condition for maximization of (ex-post) consumer surplus with respect to data collection must be satisfied by any interior maximizer (my statement focuses on this case of $d^C < d^{max}$):

$$\frac{\partial CS(d,p)}{\partial d} = \int_0^{\theta(d,p)} U_d(d,p,\theta) g(\theta) d\theta = 0 \tag{2.65}$$

Observe that by assumption $U_{d,\theta} < 0$, if U_d is weakly *positive* at some $(d, \theta(d,p))$ it is strictly positive for all $\theta < \theta(d,p)$. Thus, increasing d would benefit the marginal consumer and all infra-marginal ones, too and such a $d < d^{max}$ cannot be maximizing $CS(d,p)$ defined above. Thus, at any maximizer d^C of consumer surplus, it must be that $U_d(d^C, \theta(d^C,p), p) < 0$.[35]

For the second statement, note that via the Implicit Function Theorem, at any interior maximizer d^C:

$$\frac{\partial \theta(d,p)}{\partial d} = -\underbrace{\frac{U_d(d^C, \theta(d^C,p))}{U_\theta(d^C, \theta(d^C,p))}}_{(-)} < 0, \text{ at } d = d^C \tag{2.66}$$

where the sign of the numerator follows from the preceding argument.

[35] I do not impose that consumer-surplus is quasi-concave and make no statement about the uniqueness of d^C for given p.

2.F.1.1 Second part of Lemma

I will show that at the unique $d^*(p)$, the derivative of total consumer surplus with respect to data is *positive*. I write down the derivative of (2.64):

$$\frac{\partial CS(d,p)}{\partial d} := \int_0^{\theta(d,p)} U_d(d,p,\theta) g(\theta) d\theta \tag{2.67}$$

where via the Envelope Theorem there is no effect via participation changes. By the argument in the Proof of Lemma 16, we know that at the value $d^*(p)$, $U_d(d^*(p), p, \theta(d^*(p), p)) = 0$, i.e., the marginal consumer's utility is maximized. But then $U_{d,\theta}$ implies that $U_d > 0$ for all active consumers $\theta < \theta(d^*(p), p)$. Thus, the above partial derivative of CS with respect to data is *positive* when evaluated at $(d^*(p), p)$.

2.F.2 Proof of Lemma 10

Define participation as a function of required data collection as $D(d,p) := G(\theta(d;p))$ and look at the slope of $\Pi(d,p)$ with respect to data:

$$\begin{aligned}\frac{\partial \Pi(d,p)}{\partial d} &= \frac{\partial D(d,p)}{\partial d} r(d) + D(d)\, r'(d) \\ &= \frac{\partial D(d,p)}{\partial d}(r_0 + r_1\, d) + D(d)\, r_1\end{aligned}$$

For $r_1 = 0$ and any $r_0 > 0$, it is obvious that Π is maximized by the same d that maximizes D, i.e., participation. By the previous Lemma, we know then that at $d^F = d^*$, aggregate consumer surplus is still increasing in data. By continuity, this will continue to hold for small values of r_1.

For the second and third points, assume that there exists a *unique* maximizer of consumer surplus $d^C < d^{max}$ (note this was not necessary to prove the first point). Then, by the assumed quasi-concavity of Π in d, $d^F(p) > d^C(p)$ if and only if the above slope is positive at d^C. I prove the second point by noting that

the slope at d^C is *decreasing* in r_0, because by Lemma 9, $\frac{\partial D(d^C(p),p)}{\partial d} < 0$. Thus, if $d^F(p) > d^C(p)$ for any $r_0 \geq 0$, it is also true for $r_0 = 0$. This is true at $r_0 = 0$ when:

$$\frac{\partial D(d^C, p)}{\partial d} r_1 \, d^C + D(d) \, r_1 > 0 \iff$$
$$\varepsilon_{D,d}(d^C) > -1$$

where I define the elasticity of participation with respect to data as: $\varepsilon_{D,d}(d) := \frac{\partial D(d,p)}{\partial d} \frac{d}{D(d,p)}$. By Lemma 9, it is negative at $d = d^C(p)$.

Appendix 2.G Additional Material

2.G.1 Proof of Lemma 15

I show that when investment impedes learning, the direct effect of investment on second-period expected consumer surplus is negative, around an equilibrium of the *consumer regime*. A mirror argument shows that the direct effect is positive when investment enables learning. At an equilibrium of the consumer regime with $\{\theta_n^C, \theta_b^C, d_n^C, d_b^C\}$ second period consumer surplus is:

$$\begin{aligned} CS_2(e_1) =& \mu\big[f(e_1,\ell)CS_2(\theta_b^C, d_b^C, \ell) + (1 - f(e_1,\ell))CS_2(\theta_n^C, d_n^C, \ell)\big] \\ &+ (1-\mu)\big[f(e_1,h)CS_2(\theta_b^C, d_b^C, h) + (1 - f(e_1,h))CS_2(\theta_n^C, d_n^C, h)\big] \end{aligned}$$

where with some abuse of notation I denote by $CS_2(\theta_b^C, d_b^C, \ell)$ the aggregate consumer surplus when participation and data sharing are given by θ_b^C, d_b^C and the firm's type in period 2 is ℓ so that the probability of a breach is $f^\ell(0)$. The partial derivative of $CS_2(e_1)$, ignoring effects via changing consumer decisions, [36], is given

[36] Whose first order impact is zero in any case, starting from equilibria of the consumer regime.

by:

$$\begin{aligned}\frac{\partial CS_2}{\partial e_1} =& -\mu\, f'(e_1;\ell)\big[CS_2(\theta_n^C, d_n^C, \ell) - CS_2(\theta_b^C, d_b^C, \ell))\big] \\ & -(1-\mu)\, f'(e_1;h)\big[CS_2(\theta_n^C, d_n^C, h) - CS_2(\theta_b^C, d_b^C, h))\big] \end{aligned} \tag{2.68}$$

where the first expression in brackets is positive and the second is negative – consumer surplus is greater the closer consumers' beliefs are to the true type. Note that the first bracketed expression can be rewritten as:

$$\begin{aligned}&CS_2(\theta_n^C, d_n^C, \ell) - CS_2(\theta_b^C, d_b^C, \ell)) \\ &\qquad = \Big(CS_2(\theta_n^C, d_n^C, \ell) - CS_2(\theta_b^C, d_n^C, \ell)\Big) + \Big(CS_2(\theta_b^C, d_n^C, \ell) - CS_2(\theta_b^C, d_b^C, \ell)\Big)\end{aligned}$$

We can use this manuplutation and the Fundamental Theorem of Calculus to rewrite the slope as:

$$\begin{aligned}\frac{\partial CS_2}{\partial e_1} =& -\mu\, f'(e_1;\ell)\left[\int_{\theta_b^C}^{\theta_n^C} \frac{\partial CS_2(\theta, d_n^C, \ell)}{\partial \theta} d\theta + \int_{d_b^C}^{d_n^C} \frac{\partial CS_2(\theta_b^C, d, \ell)}{\partial d} dd\right] \\ & -(1-\mu)\, f'(e_1;h)\left[\int_{\theta_b^C}^{\theta_n^C} \frac{\partial CS_2(\theta, d_n^C, h)}{\partial \theta} d\theta + \int_{d_b^C}^{d_n^C} \frac{\partial CS_2(\theta_b^C, d, h)}{\partial d} dd\right]\end{aligned}$$

which is then rewritten as:

$$\begin{aligned}&\int_{\theta_b^C}^{\theta_n^C} (-1)\left[\mu\, f'(e_1;\ell)\frac{\partial CS_2(\theta, d_n^C, \ell)}{\partial \theta} + (1-\mu) f'(e_1, h)\frac{\partial CS_2(\theta, d_n^C, h)}{\partial \theta}\right] d\theta \\ &+ \int_{d_b^C}^{d_n^C} (-1)\left[\mu\, f'(e_1;\ell)\frac{\partial CS_2(\theta_b^C, d, \ell)}{\partial d} + (1-\mu) f'(e_1, h)\frac{\partial CS_2(\theta_b^C, d, h)}{\partial d}\right] dd\end{aligned}$$

I will show that both the top integral (A) and the bottom one (B) are *negative* when investment impedes learning. I start from integral A. I will show that the

integrand is negative for all $\theta \in [\theta_b^C, \theta_n^C]$:

$$(-1)\left[\mu\, f'(e_1;\ell)\frac{\partial CS_2(\theta, d_n^C, \ell)}{\partial \theta} + (1-\mu) f'(e_1, h)\frac{\partial CS_2(\theta, d_n^C, h)}{\partial \theta}\right] =$$

$$\underbrace{-f'(e_1,h)}_{(+)}\left[\mu\,\frac{f'(e_1;\ell)}{f'(e_1,h)}\,\underbrace{\frac{\partial CS_2(\theta, d_n^C, \ell)}{\partial \theta}}_{(+)} + (1-\mu)\frac{\partial CS_2(\theta, d_n^C, h)}{\partial \theta}\right] < \quad \text{by (2.33)}$$

$$-\,f'(e_1,h)\left[\mu\,\frac{f(e_1;\ell)}{f(e_1,h)}\frac{\partial CS_2(\theta, d_n^C, \ell)}{\partial \theta} + (1-\mu)\frac{\partial CS_2(\theta, d_n^C, h)}{\partial \theta}\right]$$

Where the positive sign of the partial derivative obtains because $\frac{\partial CS_2(\theta, d_n^C, \ell)}{\partial \theta} = g(\theta)U(\theta, d_n^C, \ell) > g(\theta)U(\theta_n^C, d_n^C, \ell)$, since that is the highest value of θ considered and $U_\theta < 0$, and $g(\theta)U(\theta_n^C, d_n^C, \ell) > g(\theta)U(\theta_n^C, d_n^C, \mu_n) = 0$ since expected utility is always increasing in the probability of facing a low-risk type.

$$\frac{-f'(e_1,h)p_1(e_1)}{f(e_1,h)}\left[\frac{\mu\, f(e_1;\ell)}{p_1(e_1)}\frac{\partial CS_2(\theta, d_n^C, \ell)}{\partial \theta} + \frac{(1-\mu)f(e_1,h)}{p_1(e_1)}\frac{\partial CS_2(\theta, d_n^C, h)}{\partial \theta}\right] =$$

$$\frac{-f'(e_1,h)p_1(e_1)}{f(e_1,h)}\left[\mu_b\frac{\partial CS_2(\theta, d_n^C, \ell)}{\partial \theta} + (1-\mu_b)\frac{\partial CS_2(\theta, d_n^C, h)}{\partial \theta}\right] =$$

$$\underbrace{\frac{-f'(e_1,h)p_1(e_1)}{f(e_1,h)}}_{(+)}\Big[\mu_b U(\theta, d_n^C, \ell) + (1-\mu_b)U(\theta, d_n^C, h)\Big]g(\theta) < 0$$

Look at the final expression in brackets:

$$\mu_b U(\theta, d_n^C, \ell) + (1-\mu_b)U(\theta, d_n^C, h) <$$

$$\mu_b U(\theta, d_b^C, \ell) + (1-\mu_b)U(\theta, d_b^C, h) <$$

$$\mu_b U(\theta_b^C, d_b^C, \ell) + (1-\mu_b)U(\theta_b^C, d_b^C, h) = 0, \forall\, \theta \geq \theta_b^C$$

The first inequality follows because under $U_{d,\theta} = 0$, d_b^C is the value that maximizes the utility of any type θ when the posterior belief is μ_b. The second inequality follows because $U_\theta < 0$. I move on to the second integral, B. Again, I will show

that the integrand is negative for all $d \in [d_b^C, d_n^C]$ and all $\theta \in [\theta_b^C, \theta_n^C]$.

$$(-1)\left[\mu\, f'(e_1;\ell)\frac{\partial CS_2(\theta_b^C, d, \ell)}{\partial d} + (1-\mu)f'(e_1,h)\frac{\partial CS_2(\theta_b^C, d, h)}{\partial d}\right] = \tag{2.69}$$

$$(-1)f'(e_1,h)\left[\mu\,\frac{f'(e_1;\ell)}{f'(e_1,h)}\underbrace{\frac{\partial CS_2(\theta_b^C, d, \ell)}{\partial d}}_{(+)} + (1-\mu)f'(e_1,h)\frac{\partial CS_2(\theta_b^C, d, h)}{\partial d}\right] \tag{2.70}$$

where the noted positive sign follows because $d < d_n^C$ is lower than the amount of data consumers would like to share if they knew they were facing a low-risk firm, hence their expected utility is still increasing at any relevant d, and $U_{d,\theta} = 0$, so this holds for all θ. Appealing again to (2.33) to swap the slope ratio of f' with the level ratio of f, we obtain that the integrand is smaller than:

$$(-1)f'(e_1,h)\left[\mu\,\frac{f(e_1;\ell)}{f(e_1,h)}\frac{\partial CS_2(\theta_b^C, d, \ell)}{\partial d} + (1-\mu)f'(e_1,h)\frac{\partial CS_2(\theta_b^C, d, h)}{\partial d}\right] =$$

$$\frac{(-1)f'(e_1,h)p_1(e_1)}{f(e_1,h)}\left[\mu_b\frac{\partial CS_2(\theta_b^C, d, \ell)}{\partial d} + (1-\mu_b)\frac{\partial CS_2(\theta_b^C, d, h)}{\partial d}\right] =$$

$$\frac{(-1)f'(e_1,h)p_1(e_1)}{f(e_1,h)}\int_0^{\theta_b^C}\left[\mu_b\frac{\partial U(\theta, d, \ell)}{\partial d} + (1-\mu_b)\frac{\partial U(\theta, d, h)}{\partial d}\right]g(\theta)d\theta$$

$$< 0$$

The sign follows because the expression in brackets is the expected marginal utility from sharing an additional unit of data when the beliefs about the firm's type is μ_b. This is precisely zero by optimality of d_b^C at those beliefs. At every other d, it is negative, and this is true for every θ in the relevant range, by $U_{d,\theta} = 0$.

3 | The Market for Ransomware Insurance

3.1 Introduction

In this chapter, I study markets for insurance against *ransomware* attacks. Ransomware is a type of malicious software which prevents firm from accessing key functions and data, usually by encrypting files. After criminal groups gain access to a company's system and infect it with ransomware, they demand a *ransom* payment - typically in untraceable cryptocurrency - in exchange for the decryption keys. If companies refuse to pay, attackers threaten to intensify the disruption to their network and to also leak or sell the company's stolen data.[1][2] The 2025 ransomware attack on Marks and Spencer in the UK caused a 16% reduction in (short-term) stock value for the retailer, whose online shop was still not restored seven weeks after the attack [The Guardian, 2025], [Financial Times, 2025].

To mitigate such risks, firms increasingly purchase cyber insurance. According to Munich Re (a major reinsurance company), ransomware attacks were the leading driver of cyber insurance claims in 2024 [Munich Re, 2025].[3] Cyber insurance policies typically cover first- and third-party liability, as well as ransom payments (see, e.g., Marsh [2025], Acronis [2025]). Additionally, and very importantly, they also offer incident response services, to help companies better deal with an ongoing

[1] These "double extortion" attacks are the norm nowadays, and have contributed to a large increase in the amount of ransom payments, see van Rooyen [2024].

[2] The UK National Crime Agency views ransomware as the largest cyber security threat, [National Cyber Security Centre, 2024], and they expect it to remain so in the future as AI allows hackers to automate their operations and reduce the cost of performing ransomware attacks.

[3] According to the same source, global premium volume in the cyber insurance market is USD 15.3 billion in 2024 and expected to more than double by 2030.

ransomware attack, as well as reputation management and regulatory compliance services, to help them deal with the aftermath.

However, it is believed that attackers can use information on victims' cyber insurance policies to determine how much ransom to demand: this view is corroborated by the study of Cong et al. [2025] who claim that the more sophisticated ransomware groups scan a victim firm's network to locate possible cyber-insurance policies before making ransom demands.[4]

This has raised concerns that the insurance sector is inadvertently contributing to higher ransom payments by insured victims, thus funding ransomware groups and incentivizing increased criminal activity. Following such concerns raised by French government officials, insurer AXA announced in 2021 that it would stop offering new policies that cover *ransom payments* in France (Euronews [2021]). More recently, the US Deputy National Security Adviser for Cyber and Emerging Technology advocated for banning insurance for ransom payments (Financial Times [2024]).

This paper investigates two main questions: first, whether the emergence of an insurance market for ransomware can reduce the welfare of firms, relative to the benchmark in which an insurance market does not exist. Second, whether and what regulatory interventions can raise welfare of firms, and potentially decrease the expected revenue of hackers. In the model I develop to answer these questions, and motivated by the above discussion, I focus on the role of (1) hackers' ability to *observe* firms' insurance contracts and (2) firms' *liquidity constraints*.

There are several reasons why I study the game under different assumptions on contract observability, not least because the prospect of hackers observing insurance contracts is generating prolific commentary by market participants and

[4]The following quote by Jamie Hart, intelligence analyst at Digital Shadows, is revealing: "They've been in the [victim's] network, they've seen it, and they're going to argue that [the victim] has coverage and they can afford to pay.." (Gooding [2024]). The title of the piece is indicative of the active debate: "Does cyber insurance increase the risk of a ransomware attack?"

policy makers. First, this allows us to identify what welfare effects are due to contract observability or lack thereof. Second, it is currently ambiguous whether contracts will be better modeled as observed once the insurance market has matured. Third, and related, studying the game under unobserved contracts allows us to understand how equilibrium will be shaped if insurers and firms adapt by e.g. not saving digital forms of the insurance policies, or adopting other measures to ensure this information does not leak. Such measures are currently recommended by cyber-security specialists.

To answer these questions I develop a theoretical model to jointly study the strategic interaction between insurance providers, firms and hackers. Firms suffer ransomware attacks with exogenous probability, and if they do, they receive a take-it-or-leave-it ransom demand by the adversary. Firms can either accept and pay the ransom, or reject the offer and suffer *business interruption*. The adversary's threat to induce business interruption is what incentivizes the firm to pay. In the event of rejection, the adversary sells the data stolen by the firm and makes positive revenue. Data is not sold if the firm pays the ransom.[5]

In the insurance market, I abstract from moral hazard and private information between insurance buyers and sellers, to focus on the novel strategic aspect of ransomware insurance provision. The insurance contract is comprised of three components: the premium, coverage for ransom payments and coverage for business interruption that firms suffer following *rejection* of the ransom demand.

How the insurance policy of firms affects ransom bargaining depends on whether hackers can observe victim's policies. In the case of *observed* contracts, I assume that upon a successful attack, the hacker learns whether the victim is insured and the precise contract it holds. Instead, in the case of *unobserved* contracts, the

[5]I assume that the adversary's promise to not leak data and not induce business interruption after ransom is paid is credible. This is consistent with observed practice and is due to unmodeled reputational concerns of the ransomware gangs. I further elaborate on this after I present the model.

hacker obtains no additional information about the firm's insurance status. Of course, even under unobserved contracts, hackers' equilibrium conjectures about the insurance market must be correct.

I begin with the analysis of **observed** contracts. The insurance contract determines the maximum ransom that a firm is willing to accept, and if that maximum acceptable ransom falls below the adversary's outside option, there is ransom rejection by insured firms. When the interruption to the firm, b, i.e. the threat, is worth more than the adversary's outside revenue, s, the profit-maximizing insurance contract induces *payment of ransom* in equilibrium.[6]

Observed contracts have *commitment value* to the firm: the equilibrium contract provides full insurance to firms against business interruption, which depresses ransom demands to hackers' disagreement payoff, s. This is the lowest ransom possible such that adversaries do not inflict severe business interruption to the firm. In equilibrium, firms are also fully insured against that ransom payment.

Adversaries are *worse off* relative to the benchmark in which insurance markets do not exist, and under monopoly, firms are equally well off. Because the equilibrium insurance contract minimizes the harm faced by insured firms, it is also the contract offered in equilibrium under perfect competition of insurers. Insurance buyers are thus made *strictly better off* under competitive insurance markets relative to the benchmark of no insurance market. In additional extensions, I show how these results remain robust to the determination of ransom via Nash Bargaining and to endogenizing the participation of adversaries.

Under **unobserved** contracts, the insurer cannot directly affect the ransom demanded from insured firms; hackers instead choose ransom demand to best respond to their conjectured insurance policy. The *strategic complementarity* between ransom demands and insurance for ransom payments leads to a multiplicity

[6]This is the natural case to consider. Otherwise, in equilibrium without insurance we would never observe ransom negotiations. Instead, adversaries would simply monetize the stolen data.

of equilibria (Proposition 15). Nevertheless, it remains true that in every equilibrium, insurance contracts maintain some of their commitment value and insured firms pay *less* for ransom than in the benchmark.

Interestingly, and in contrast to the monopoly case of observed contracts, under *any* equilibrium of Proposition 15, firms are *better off* than in the benchmark of no active insurer. Under unobserved contracts, uninsured firms off the equilibrium path benefit from a positive externality, and face lower ransom than if attackers knew they were facing an uninsured firm. Thus, the monopolist can extract less surplus in the insurance market and firms are better off.

But even though in any equilibrium firms are made better off by the presence of a monopoly insurer, multiplicity creates scope for regulatory intervention to ensure that the lowest-ransom equilibrium of Proposition 15 is being played. Just as in the unique equilibrium under observed contracts, the lowest equilibrium ransom and corresponding ransom insurance equal hackers' outside profit s. That equilibrium is simultaneously the firm-best and adversary-worst.

Finally, I consider a setting with **liquidity constraints**. I assume that the level of liquidity is known by hackers and thus in the absence of insurance markets, firms pay ransom equal to their liquidity.[7] In equilibria with an insurance market, liquidity constraints do not bind for insured firms, and under *unobserved* contracts, the multiplicity of Proposition 15 arises again.

However, unlike the case without liquidity constraints, there are now equilibria in which adversaries become *better off* relative to the benchmark. Intuitively, this is when the commitment value of insurance contracts is not enough to compensate for the relaxation of firms' liquidity constraints, and insured firms face higher ransom demands than in the benchmark. In such equilibria, firms are made worse off by the existence of an insurance market. The policy prescription of the previous

[7] I assume that liquidity is sufficient to compensate hackers for s.

section remains valid, but an additional one emerges: to guarantee equilibria in which adversaries are worse off, regulators should cap insurance coverage for ransom payments below firms' standalone liquidity levels.

In the next section, I review related literature. In Section 3.3, I present the model with observed contracts. I discuss the main modeling assumptions, derive the model's equilibrium and present extensions. In Section 3.5, I move on to the model with unobserved contracts, and finally, in Section 3.6, I extend the model to account for liquidity constraints, under both observed and unobserved contracts. I derive the welfare and policy implications, and then conclude.

3.2 Related Literature

The work closest to mine is Balasubramanian [2021], who also studies insurance with strategic ransomware attackers. In his model, ransomware insurance offers risk-sharing value but by raising firms' liquidity always leads to higher equilibrium ransom demands and more active hackers. However, in contrast to my model, insurance coverage does not discriminate between business-interruption and ransom payments, rather a uniform coverage amount is specified. This implies that insurance has no strategic commitment value in his model, and only relaxes firms' liquidity constraints.

Theoretically, my work contributes to the vast literature on the economics of insurance, but I abstract from well-studied aspects of moral hazard (see survey by Parra and Winter [2025]) and private information. I do so to focus on the novel role of ransomware insurance in manipulating the bargaining equilibrium between hackers and victim firms, and thus influencing the level of harm that firms face. In studying the role of contract observability, my work relates to the seminal contributions of Katz [1991] and Bolton and Scharfstein [1990]. The work of Katz [1991] asks whether unobservable contracts between a principal (insurer,

in my case) and an agent (firm) can serve as pre-commitments for the agent in games played against other agents (hackers). In a financial contracting application, Bolton and Scharfstein [1990] show that under publicly observed contracts, a lender's optimal financing contract is chosen to influence the product-market game and protect the borrower against predation by an incumbent. However, offering that same contract is not credible under unobserved contracts, hence predation occurs in equilibrium.

Ahnert et al. [2022a] and August et al. [2025] share with my paper that hackers choose whether to earn revenue via a "traditional" data breach, i.e. via the sale of stolen firm data, or via ransomware. The work of Ahnert et al. [2022a] studies an environment with unobserved investments in security. Hackers choose to shift all losses to *customers* so that in equilibrium firms do not invest at all in cyber security. If a market for security certifications (security signals) exists, hackers endogenously switch to ransomware business models to shift losses away from consumers and dissuade firms from purchasing these signals. The emergence of ransomware shifts the incidence of cyber-attack losses from consumers to firms and can induce higher levels of security relative to an environment with only traditional data breaches. In August et al. [2025], hackers with heterogeneous entry costs sort between traditional breaches and ransomware attacks. The authors ask how social welfare changes once the availability of cryptocurrency makes ransomware attacks feasible, relative to a world with only traditional breaches. The option to pay ransom can potentially benefit firms if equilibrium ransom is small relative to the harm they suffer from a traditional data breach.

The recent contribution of Cartwright et al. [2023] discusses a variety of channels through which cyber insurance can affect the level of ransom paid. They explicitly recognize that in contrast to the moral hazard considerations and liquidity effects that point towards higher ransom paid, insurance that covers *incident response*

and ransom-rejection costs should lower the incentive to pay ransom. My model is the first work I am aware of that explicitly models the separate components of insurance for ransomware attacks. Meurs et al. [2023] is a first attempt at *empirically* understanding whether firms that hold cyber insurance policies are attacked more frequently and pay higher ransom, using data from self-reported ransomware attacks in the Netherlands. The paper by Cong et al. [2025] studies the transcripts of 700 ransomware negotiations and offers a thorough description of how the main ransomware gangs operate.[8]

More broadly on the topic of cyber-insurance, the work of Böhme and Schwartz [2010] models the impact of cyber-insurance, and focuses on aspects of cyber-security like correlated risk and security externalities between insured firms, but predates the advent of ransomware. Laszka et al. [2017] offers a model of strategic investment in backup and ransom negotiations between firms and adversaries. Cartwright and Cartwright [2019] deals with the interesting question of reputation formation by long-lived ransomware gangs, who after every successful attack choose whether to return the firm's data or not. The main finding is that equilibria with reputational incentives are more likely to arise when there is a small number of well-known adversaries, rather a mass of potential entrants.

3.3 Baseline Model

The game takes place over a single period comprising of several stages. There is a single insurance provider, a unit-mass of identical *firms*, which operate in an industry subject to ransomware attacks and a unit mass of identical *adversaries*,[9] hackers who engage in ransomware attacks. The firms maximize expected utility and are risk-averse with concave Bernoulli utility function over end-of-

[8]Most interestingly in relation to my work, it discusses how the major gangs resemble proper businesses and how they look for information about the assets and insurance policies of firms they attack, in order to optimize ransom offers.

[9]For most of the following analysis, we can also consider a single firm and a single adversary.

period wealth levels that satisfies $u' > 0, u'' < 0$. The adversaries and insurer are risk-neutral profit maximizers. The insurer plays first by choosing an insurance contract with terms $(p, M_r, M_b) \in R^3_+$. Aside from the premium, p, the insurance contract is comprised of (weakly positive) terms M_r and M_b, whose function I explain below. Then, each firm has the option of paying the premium to purchase the insurance contract. After a firm has made an insurance-purchase decision, it operates in the market and is "matched" with an adversary who attempts to breach its network. Each adversary is successful with *exogenous* probability $q \in (0, 1)$. I assume that successes occur independently of whether insurance has been purchased or not. In addition, the probability q is known to insurer and firm at the time the insurance contract is signed.

If an attack is unsuccessful, which occurs with probability $(1 - q)$, the game ends. If the attack is successful, the firm and corresponding adversary enter the bargaining stage of the game.

Bargaining Subgame. Given the information available to him at the beginning of the bargaining subgame, the adversary makes a take-it-or-leave-it (TIOLI) ransom offer r to the victim firm. I will analyze two regimes separately: first, a regime of (ex-post) observed contracts, in which once an adversary breaches a firm's network, he immediately learns whether the victim firm has signed an insurance contract and the precise contract terms M_r and M_b. Second, a regime of *unobserved* contracts, in which the adversary observes *neither* whether he has breached the network of an insured or uninsured firm *nor* observes the contract offered by the insurer.

If bargaining is successful and the firm accepts to pay ransom r, the adversary's payoff is r. This is regardless of whether he is bargaining with an insured or uninsured firm, and of whether contracts are observed or not. On the other hand, if there is bargaining breakdown, the adversary earns his outside option of s, the

payoff from monetizing information stolen by the firm. In the event of successful bargaining firms suffer $r+b^{low}$, which is the ransom plus business interruption (BI) damage. Parameter b^{low} is the business interruption damage caused to the firm upon the occurrence of a breach, regardless of bargaining outcome. Without loss, I assume $b^{low} = 0$. This is the part of BI costs that is *sunk* at the time of bargaining, hence a positive value would not qualitatively impact any of my results.[10] In the event of bargaining breakdown, the firm suffers harm b^{high}, and since I focus on $b^{low} = 0$, I will omit the superscript and refer to b^{high} as b. On top of those payoffs, if the firm has purchased an insurance contract with terms M_r, M_b, it receives M_r compensation if it pays the ransom, and M_b compensation if it rejects the offer. The insurance contract does not directly affect the adversary's payoffs, but may do so indirectly via affecting whether the firm will accept a given ransom demand r in equilibrium, and the value of the equilibrium demand r. Before moving on to the equilibrium analysis, I briefly discuss the modelling of payoffs presented above.

3.3.1 Discussion of modelling and assumptions

Interpretation of b. In reality, the harm that adversaries can cause to the firm following rejection of the ransom offer has various components: (1) business interruption induced by hackers not releasing data and/or network resources that are necessary for the firm's operations, (2) exposure to regulatory punishment (e.g., GDPR fines) if data leaked (see this recent example, Woodruff Sawyer [2024]), (3) reputational harm if consumers become aware of data sale.[11] The recently observed occurrence of "double extortion" means precisely that hackers can threaten to cause both business interruption and data theft in the event of

[10] This does not mean that I assume b^{low} is small in absolute size or relative to any of the other parameters in the model. In fact, as August et al. [2025] discuss, even when ransom is paid, there are significant additional costs from a ransomware attack.

[11] Current data-protection regulation requires consumers to be notified only if their data is leaked, but investors may have to be informed even in the absence of personal data leakage.

bargaining breakdown (see Cartwright et al. [2023], Cong et al. [2025], Böhme and Schwartz [2010]). All components of the harm caused to the firm following bargaining breakdown are increasing in the "size" of the breach, which in this model is represented by the scalar b.

Interpretation of M_b Varying degrees of insurance can be provided against all the aforementioned components of the threat: the insurer can provide (1) technical remedies and support via specialized IT staff to mitigate business interruption, (2) legal support to reduce regulatory penalty, (3) crisis management support and credit monitoring for consumers. These are indeed some of the key components of modern cyber insurance policies (see also Ransomware Task Force, Acronis [2025], Marsh [2025]). For example, the Financial Times reported after the recent data breach of $M\&S$ that their cyber insurance policy "..would cover both first-party losses, such as lost sales and incident response costs, as well as third-party losses, such as legal liabilities related to the data breach..".[12] I abstract away from specifying the means of compensation to firms that suffer business interruption and only refer to payment M_b, the insurance payment in the event of ransom rejection.

Symmetric information between adversary and victim. As Cong et al. [2025] find, one of the first pieces of information that ransomware attackers look for in a firm's network is the financial statement, which gives them detailed knowledge of the firm's assets and profitability. The hackers use estimates of revenue lost as a result of encrypted resources to infer the value of b.

Insurance does not affect attack probability. Regarding insurance, I assume that adversaries (1) get matched to insured and uninsured firms with the same probability, i.e. there is no ex-ante targeting of either type of firm and (2) and insurance does not offer better security, i.e. the probability of a successful attack

[12] https://www.ft.com/content/723b6195-1ce7-4b5f-94f5-729e9152c578, accessed July 24, 2025.

is q for both insured and uninsured firms alike. It is ex-ante ambiguous how insurance could affect the total probability of a successful ransomware attack; on one hand, if hackers anticipate larger payments by firms with insurance (holding other firm characteristics fixed), this would induce them to find out which firms are insured and target them. On the other hand, pooling cyber-security expertise means cyber insurers can provide advice to their customers on how to avoid ransomware attacks, which is a common function of insurance in many other contexts. I abstract away from these considerations by assuming q remains independent of the decision to get insured.

Credibility of adversaries. Finally, it is important that the firm does not suffer harm b if it pays the ransom offer. The underlying assumption is that the ransomware group credibly commits not to further harm the firm if the ransom is paid, which is plausible due to reputational concerns of the large ransomware groups that dominate this space. The analysis of ransom negotiation transcripts by Cong et al. [2025] offers strong evidence of such reputational concerns.

3.3.2 Equilibrium with observed contracts

Bargaining subgame

First, I derive a firm's best response to a ransom offer r. An uninsured firm will accept a ransom offer r if and only if $r \leq b$. An insured firm with insurance terms M_r, M_b accept a ransom offer r if and only if $r - M_r \leq b - M_b$, i.e. if the net harm from accepting the ransom demand is lower than the net harm of rejecting the offer. In the case of observed contracts, the risk-neutral adversary who faces an uninsured firm will set a TIOLI ransom offer $r(\emptyset) = \max\{b, s\}$, and in the case of an insured firm will make an offer of $r(M_r, M_b) := \max\{b + M_r - M_b, s\}$. The adversary finds it worth it to extract ransom if $r(M_r, M_b) > s$. If the maximal ransom the firm is willing to pay is not enough to compensate for the value of

selling the data, the adversary will ask for a high enough ransom that the firm will reject. I summarize the above in the following Lemma:

Lemma 17. *When facing a firm with insurance contract* (M_r, M_b)*, a best-response*[13] *for the adversary is to make a TIOLI offer* $BR^A(M_r, M_b) = r(M_r, M_b) = \max\{b + M_r - M_b, s\}$*. This remains true for the case of an uninsured firm, for which* $M_r = M_b = 0$*.*

Since the attacker has full bargaining power relative to the firm, when the firm accepts the ransom it is indifferent between acceptance and rejection, which yields harm $(b - M_b)$. Thus, the firm faces harm $(b - M_b)$ regardless of whether it accepts or rejects the offer $r(M_r, M_b)$. The payoff of the hacker, however, is different in the two cases.

Proposition 10. *Given an insurance contract* (M_r, M_b)*: the following is true in any subgame equilibrium*[14] *of the bargaining game that begins after the successful breach of a firm with insurance contract* (M_r, M_b)*.*

1. *If* $r(M_r, M_b) > b - M_b + M_r \iff s > b - M_b + M_r$*, the firm finds it strictly optimal to reject. The adversary earns* s *and the firm suffers harm net of insurance payments* $(b - M_b)$*.*

2. *If* $r(M_r, M_b) = b - M_b + M_r \iff s \leq b - M_b + M_r$*, the firm is indifferent between accepting and rejecting the offer. Regardless of decision, the harm net of insurance payments is equal to* $b - M_b$*.*

Thus, for any insurance contract, the insured firm's equilibrium monetary loss when suffering a breach is $(b - M_b)$*. For an uninsured firm, it is* b*.*

By the above result, the expected utility of an insured firm does not directly

[13] For $r(M_r, M_b) > s$, the best response is unique. In general, the best-response is to ask for any ransom $r \geq r(M_r, M_b)$ when $r(M_r, M_b) = s$.

[14] There exists a multiplicity of subgame equilibria that only differ in the hacker's ransom demand in the case $r(M_r, M_b) < s$ in which the hacker wants the demand to be rejected. Across every one of these subgame equilibria, payoffs of both parties are constant.

depend on M_r directly, since the adversary fully adjusts the ransom demand to extract M_r:

$$U^I(p, M_r, M_b) = (1-q)\,u(w-p) + q\,u(w-p-(b-M_b)) \tag{3.1}$$

The expected utility of an uninsured firm is:

$$U^N = (1-q)\,u(w) + q\,u(w-b) \tag{3.2}$$

Taking into account how the choice of insurance contract shapes equilibrium outcomes in the bargaining subgame, the monopolist insurer selects the profit-maximizing contract to offer.

3.3.3 Monopolist insurer

I assume there is symmetric information between firms and insurer and that the insurer makes a TIOLI offer of an insurance *contract* to firms. The risk-neutral monopolist insurer chooses a contract comprising of a premium p and terms M_r, M_b to maximize revenue net of insurance payments, subject to the firm being indifferent between buying insurance and not doing so.

$$\max_{p, M_r, M_b} \mathbb{E}\Pi(M_r, M_b, p) = p - \mathbb{E}\Pi(\text{insurance payment})$$

$$\text{subject to: } IC^{\text{hacker}}, IC^{\text{firm}}, IR^{\text{firm}}$$

The IC constraints of firm and hacker refer to the strategies of those players in the *bargaining* subgame, whereas the IR constraint of the firm refers to its incentive to buy insurance relative to the outside option of staying uninsured. According to Proposition 10, the expected profit function $\mathbb{E}\Pi(M_r, M_b, p)$ is discontinuous at points where the victim firm's equilibrium decision switches from "accept" to

"reject" and thus the insurance payment changes between M_b (insurance payment following rejection) and M_r (insurance payment following acceptance). We look for the most profitable contract that induces payment of ransom and the most profitable contract that induces rejection of the ransom offer.

Profit-maximizing contract that induces ransom payment.

In order to have ransom-payment in equilibrium, the relevant incentive-compatibility constraint of the adversary that must be satisfied is $M_r - M_b + b \geq s$; if that is true, the adversary makes a ransom offer that an insured firm accepts. Under this IC constraint, the expected payment the insurer makes to an insured firm is $q\,M_r$. For given choices of M_r, M_b, the insurer sets the premium p^* to extract all of the firm's expected surplus, and the constrained maximization problem is:

$$\max_{p,M_r,M_b} \{p - qM_r\}, \quad \text{s.t.} \quad s \leq b + M_r - M_b \quad \text{and} \quad U^N \leq U^I(p, M_r, M_b) \quad (3.3)$$

Both constraints will bind at the optimal solution. The firm's IR constraint must bind in equilibrium, otherwise the insurer would raise the premium, without affecting any of the IC constraints. Next, I argue that the adversary's IC constraint must also bind in equilibrium, i.e., the adversary must be made indifferent between making revenue via the ransom offer or via the outside option of selling data. If constraint were slack, the insurer would increase M_b, lowering the ransom demand that insured firms face. Thus, he could raise the premium extracted by firms, and since M_b is not paid in an equilibrium with ransom payments, the expected cost of insurance would not increase in the process.[15] The lack of moral hazard and private information implies that offering full-insurance is the insurer's best response, which in an equilibrium with ransom payments requires that $M_r = r$. These two conditions, imply that $M_b^* = b$ and $M_r^* = s$, with the premium set to extract

[15] Alternatively, we observe that by Proposition 10 the firm's loss is always $(b - M_b)$, hence symmetric information between firm and insurer and lack of moral hazard imply that full-insurance is profit-maximizing: $M_b^* = b$.

firms' surplus relative to remaining uninsured, $u(w - p^*) = U^N$.

Proposition 11. *The profit-maximizing contract that induces an equilibrium in which the firm pays ransom has* $(M_r^* = s, M_b^* = b)$ *and premium that is the unique solution to* $U^I(p^*, s, b) = U^N$. *The expected profit of the insurer is* $p^* - q\,s$.

Profit-maximizing contract that induces sale of data.

In this case, the relevant incentive compatibility constraint for the adversary is $s \geq M_r + b - M_b$. The insurer pays $q\,M_b$ in expectation. The constrained maximization problem is:

$$\max_{p, M_r, M_b} \{p - qM_b\}, \quad \text{s.t.} \quad s \geq b + M_r - M_b \quad \text{and} \quad U^N \leq U^I(p, M_r, M_b) \tag{3.4}$$

Proposition 10 still holds, and the net harm of insured firms is $b - M_b$. Full insurance against the relevant harm is again optimal, $M_b^* = b$, thus the adversary's IC constraint becomes $M_r \leq s$. *Any* choice of $M_r^* \in [0, s]$ solves the constrained maximization problem of the insurer, i.e. the IC constraint is now *slack*. Since M_r affects neither the utility of the insured firm (by Proposition 10) *nor* the expected insurance payment, it can be freely adjusted to satisfy the IC constraint for any value of $s \geq 0$. The optimal premium solves $U^I(M_b^*, M_r^*, p^*) = U^N \iff u(w - p^*) = U^N$. When any of the profit-maximizing contracts $(p^*,\ M_b^* = b,\ M_r^* \leq s)$ is used, the insurer makes expected profit $p^* - qb$.

Comparing profits

To compare profits between the two candidate optimal contracts, first note that between the two cases, p^* will be the same, equal to $w - u^{-1}(U^N)$. This is because both contracts offer full insurance to the insured firm and the value of the outside option of the firm, U^N, is the same across the two scenarios. The comparison of profits then simply boils down to a comparison of expected insurance payouts, $q\,s$

and $q\,b$.

Proposition 12. *In the game with a monopolist insurer and observed contracts:*

- *If $b > s$, in the unique SPNE, the insurer offers $M_b = b$, $M_r = s$, p^*. In equilibrium all firms purchase insurance. The adversary makes offer $r(s, b) = s$ to an insured firm, and firms accept the ransom offer. Off-path, adversaries make ransom offer $r(0, 0) = b$ to an uninsured firm, and that offer is accepted.*

- *If $b < s$, there is a continuum of payoff-equivalent SPNE, which on the equilibrium path[16] differ only in the value of $M_r^* \in [0, s]$. In every equilibrium, all firms purchase insurance. The adversary makes offer $r(M_r, b) = s$ to an insured firm, and firms reject the ransom offer. Off-path, adversaries make ransom offer $r(0, 0) = s$ to an uninsured firm, and that offer is rejected. Across every equilibrium, premium p^* is the same.*

In any equilibrium, p^* is set to make the firm indifferent between purchasing insurance and staying uninsured. In any equilibrium, adversaries earn expected payoff of qs.

Under $b > s$, we can think of the two contract terms M_b, M_r as having distinct roles in shaping equilibrium: $M_b = b$ provides *commitment value* to the firm in its bargaining and depresses the ransom demand it faces. Coverage for the ransom payment $M_r = s$ provides the firm with full insurance against the ransom it pays, i.e. offers *risk-sharing* value. An even lower value of M_r would induce the adversary to take advantage of their outside option, so the monopolist is effectively "bribing" the adversary just enough to not cause harm b to the firm. The equilibrium contract is the **unique** contract that simultaneously achieves three objectives: (1) offers full insurance to the firm, (2) ensures costly business

[16] If $b < s$, equilibria can also differ in the ransom demand, as explained in the footnote of Proposition 10, but the payoffs are constant across equilibria.

interruption is avoided, and (3) minimizes the attacker's ransom demand, subject to (1) and (2).

Under $s > b$, insurance contracts only offer risk-sharing value to firms, since in that case there is mutually acceptable ransom payment and no strategic interaction in the bargaining stage.[17]

Benchmark: No insurance market. Consider the benchmark equilibrium in which there is no active insurance provider, and all active firms are successfully breached with probability q. I denote the expected utility of firms in the benchmark by U^0 and the corresponding hacker payoff by π^0. Maintaining the TIOLI assumption for hacker ransom demands yields the following:

Lemma 18 (No insurance market)**.** *Without an active insurance market there is a unique equilibrium. Hackers earn expected payoff* $\pi^0 = q \max\{s, b\}$ *and firms' expected utility is* $U^0 = q\,u(w-b) + (1-q)\,u(w)$.

At this point we can ask how equilibrium changes relative to the case in which the insurance market does not exist, in particular, how the welfare of firms and adversaries change relative to π^0, U^0. Answering that question will be key to addressing policy concerns about the existence of markets for ransomware insurance.

The equilibrium expected utility of firms in the absence of an insurance market, U^0 is the same as the equilibrium expected utility of uninsured firms in the game with a monopolist insurer, U^N. This is a direct consequence of the observed contracts assumption.[18] In equilibrium of Proposition 12, firms are indifferent between buying insurance and not doing so, $U^* = U^I = U^N$. Thus, firms' welfare is unaffected by the entry of an insurance monopolist with who fully extracts

[17]Note that the contract $M_r = s, M_b = b,\ p^*$ can be supported in equilibrium in either case, but the (insured) firm's equilibrium strategy differs between the cases $b > s$ and $b < s$. When $b < s$, there can be no equilibrium in which this contract is offered and the firm accepts the ransom offer. The insurer would profitably deviate to instead offer $M_r = s - \epsilon$, inducing rejection and thus reducing the insurance payment without changing p^*.

[18]And as we shall see later, of the assumption that there is no endogenous margin of hacker participation.

their equilibrium surplus. On the other hand, adversaries earn revenue $\pi^{\text{obs}} = qs \leq \pi^0 = q \max\{s, b\}$ and are thus (weakly) worse-off in equilibrium relative to the absence of an insurance market.

Corollary 2. *Under the presence of the monopolistic insurer, the adversary's expected equilibrium payoff is $\pi^{obs} = qs$, in any equilibrium. The presence of the monopolistic insurer strictly reduces the hacker's payoff relative to the case in which insurance provision is unavailable if $b > s$ and leaves it unchanged otherwise. Firms' welfare is unaffected by the presence of the monopolist, $U^{obs} = U^0$.*

Empirical Predictions The above results generate empirical predictions. First, the presence of the insurance provider does not change whether ransom is paid in equilibrium, or not. When the business interruption from bargaining failure is large, then the contract inducing bargaining failure is relatively less profitable, and the insurer prefers to induce ransom payment in equilibrium, so long as the hacker's outside option from selling the data is not too large. We should observe ransom payment in equilibrium precisely when $b > s$, and this is the same without an active insurance market. Second, and similarly, observe that in the insurance equilibrium, the occurrence of ransom transfers is not affected by whether firms are insured or not. Off the equilibrium path, uninsured firms pay ransom if and only if $b > s$, and the same is true for insured ones. Thus, we should not observe different rates of ransom payment across insured and uninsured firms.[19] Third, however, the ransom payments made by uninsured firms are *higher*, and similar to those of firms in markets without active insurance providers. Fourth, when ransomware attacks do happen in the absence of an insurance market, i.e. when $b > s$, the insurance market does offer insurance for ransom payments in equilibrium, and the equilibrium value of M_r should be close to the value of the adversaries' outside

[19]There are no uninsured firms in equilibrium; this prediction would be generated by a simple extension of the model in which a fraction of (otherwise identical) firms simply never considers insurance, a natural assumption to make in a nascent market such as cyber and ransomware insurance.

option from bargaining breakdown.

These empirical predictions are based on Proposition 12, which relies on a model with ex-post observed contracts. Before analyzing the case in which adversaries do not observe the firm's insurance contract, I examine the robustness of these results to some important extensions of the observed-contracts baseline.

3.4 Extensions

3.4.1 Ransom determined via Nash Bargaining

I first show that the results extend to a relaxation of the TIOLI offer assumption in the ransom bargaining game. I assume instead that the equilibrium ransom is determined via Nash Bargaining, with bargaining power of the adversary given by $\beta \leq 1$. If ransom is paid in equilibrium, the value is given by the Nash Bargaining Solution $r^{NB}(M_r, M_b) = (1-\beta)s + \beta(M_r - M_b + b)$. The baseline corresponds to $\beta = 1$. As in the baseline, it remains true for any β that ransom is paid in the bargaining stage if and only if it is the efficient outcome (given the insurance contract), i.e., if $M_r - M_b + b \geq s \iff r^{NB}(M_r, M_b) \geq s$.

Proposition 13. *Fix $b > s$: for any $\beta \in (0,1)$, all firms purchase insurance in the unique equilibrium and the equilibrium contract involves $M_r^* = s, M_b^* = b$ and p^* such that the insurer extracts all surplus of firms. For any β, the adversary's equilibrium payoff in the bargaining stage is equal to their outside option, s. The equilibrium premium is increasing in β.*

As in Proposition 12, to maximize profits subject to inducing ransom payment, the insurer wants to reduce the ransom paid by insured firms and offer insurance against that ransom payment. For any $\beta \in (0,1)$, the unique M_r, M_b pair that minimizes the ransom and also provides full insurance is $M_r^* = s, M_b^* = b$. The equilibrium premium is increasing in β, because the ransom paid by uninsured

firms off the equilibrium path is increasing in hackers' bargaining power.

3.4.2 Competitive insurance market

I explore how the introduction of (perfect) competition in the insurance market changes equilibrium outcomes. I assume that, instead of a monopolist, a continuum of identical insurance providers operate in the market, and they simultaneously post their contracts at the beginning of the game. The insurers are not capacity constrained, hence all firms purchase insurance from the seller whose contract offer yields the highest expected utility. In equilibrium, each insurer will offer the unique contract that maximizes insured firms' expected utility, subject to the break-even constraint.

The premium charged to firms will be different relative to the case of a monopolist insurer, but the equilibrium M_r, M_b terms will not. Focusing again on the case of $b > s$, in Proposition 12, the monopolist insurer offers the terms (M_r, M_b) that maximizes the customer's expected utility, by minimizing the payment to the adversary and also offering full insurance.[20] Thus, the same (M_r, M_b) is also offered in competitive equilibrium with homogeneous insurers. The difference is that the premia charged will equal expected insurance payments, so that insurers just break even.

Proposition 14. *If $b > s$, the game with competition between identical insurance firms has a unique subgame perfect equilibrium in symmetric strategies. All insurers set $M_r^* = s, M_b^* = b, p^* = qs$. All firms buy insurance and accept the ransom offer $r(s, b) = s$. Off-path, adversaries make offer $r(0, 0) = b$ to uninsured firms, who accept the offer.*

In the unique equilibrium, insurers in the perfectly competitive market are only compensated for the marginal cost of offering an additional insurance contract,

[20]Note that the (M_r, M_b) pair that maximizes the expected utility of insured firms does not depend on the insurance premium charged.

$q\,M_r$. Since M_b is never paid out in equilibrium, the commitment-value of insurance is supplied at zero marginal cost. This implies that in the perfectly competitive equilibrium, risk-neutral insurers are not compensated for the *commitment* value of insurance, but only for the risk-sharing component of value. For that reason, firms are strictly better off relative to the benchmark equilibrium without an insurer present, i.e. $U^* > U^0$.

3.4.3 Endogenous participation of adversaries

We can also verify that the equilibrium insurance terms (M_r^*, M_b^*) of Proposition 12 remain the ones offered in equilibrium if we extend the model to allow for endogenous participation of adversaries. Assume that each one among the continuum of adversaries has a participation cost c drawn from cdf F and that adversaries participate if and only if their expected revenue exceeds their entry cost. The key assumption I make in this extension is that adversaries do not observe the true contract offered to firms at the time they make their participation decision, rather only at the time they make their ransom demand. As discussed in my introduction, I assume hackers obtain this information by gaining access to a particular firm's network and files, which is consistent with individuals not having this information at the time they decide to become ransomware hackers.[21]

For given participation of adversaries $m \leq 1$, each firm is breached with probability $q\,m$ and according to Proposition 10 suffers net harm $(b - M_b)$. Since by assumption, the insurer cannot use the insurance contract to manipulate *participation* decisions of adversaries, by the same logic of Proposition 12, the unique best response (if $b > s$) is to offer $M_r = s$, $M_b = b$. Understanding this, adversaries know their expected revenue from being active is $\pi = q\,s$ and we obtain the following:

[21]This is also consistent with my assumption that firms cannot be targeted on the basis of having purchased insurance.

Lemma 19. *If $b > s$, there is a unique equilibrium in the game with endogenous participation of adversaries. In equilibrium, ransom offers to insured firms are $r^* = s$ and firms accept. Expected revenue of active adversaries is $\pi^* = q\,s$ and participation given by $m^* = F(\pi^*)$. The contract offered is ($M_r^* = s$, $M_b^* = b$) and p^* that makes firms indifferent between buying insurance and not doing so.*

Since this is the contract that minimizes the expected payoff of each active hacker, it is also the contract that *minimizes participation*. Thus, under this new set of assumptions, there exists an additional margin by which the presence of an insurer will reduce adversarial activity relative to the benchmark. And it has another implication: the expected utility of uninsured firms is greater than $U^0 = (1 - qF(\pi^0))u(w) + qF(\pi^0)u(w - b)$, since the mass of active hackers is smaller in the insurance equilibrium. Since in equilibrium $U^I = U^N$, it follows that insured firms are strictly better off relative to the benchmark, even if the monopolist insurer has full bargaining power vis-a-vis insurance buyers.

3.5 Unobserved Contracts

In this section, I explore the extent to which the results of Proposition 12 are sensitive to the assumption that adversaries can observe the precise insurance contract a breached firm has signed. In particular, I assume that the adversary cannot observe the contract terms offered by the monopolist in the first stage of the game and cannot observe whether the breached firm has signed a contract or not. The risk-neutral adversary makes a TIOLI ransom offer to the victim once the breach has been successful, balancing the payoff from acceptance to the probability of rejection. If that offer is rejected, the adversary earns the outside option of s. In the main text, I present the main arguments for finding pure-strategy equilibria and defer presentation of the complete argument to the Appendix.

Derive the adversary's best-response.

Given the (single) contract offered M_r, M_b, p, an adversary understands that an insured company will accept the ransom offer r if and only if $r - M_r < b - M_b$ and an uninsured company will accept if $r < b$. Given anticipated fraction of insured firms μ, the adversary perceives probability of ransom offer r being accepted equal to $P(r) := \mu\ 1\{r - M_r < b - M_b\} + (1 - \mu)1\{r \leq b\}$ sets ransom to maximize $\{P(r)\, r + (1 - P(r))s\}$. I restrict my attention to pure-strategy equilibria in which the insurer offers the same contract terms to all companies and all companies choose to buy it. The latter is without loss, since all companies will buy insurance in equilibrium. Given this, $P(r) = 1\{r - M_r < b - M_b\}$ and the adversary sets:

$$BR^A(M_r, M_b; \mu = 1) = \begin{cases} r : r \geq s, & \text{if } s > r^{max}(M_r, M_b) \\ r^{max}(M_r, M_b), & \text{if } s < r^{max}(M_r, M_b) \end{cases}$$

where $r^{max}(M_r, M_b) := b + M_r - M_b$ is the maximum ransom that an insured firm is willing to pay, given it has signed a contract with terms M_r, M_b.[22] Again, the best-response of the adversary is not uniquely pinned down if he wants to induce rejection of the offer.

Derive insurance company's best response.

The insurer does not directly affect the adversary's offered ransom with the choices of contract terms, and firms cannot not affect the adversary's conjecture on the mass of firms that get insured. Additionally, the insurer cannot directly affect the firms' *outside option* of not getting insurance via its choice of M_r, M_b, however that outside option will depend on the contract term in equilibrium through the adversary's choice of ransom offer.[23] For given ransom demand $r \geq s$, the insurer must again decide whether to set contract terms that induce acceptance or rejection of this offer by insured firms. I focus on the case of $r \leq b$, and I show in the Appendix, that the insurer will want to induce payment of ransom. In that case,

[22] In other words, a best response of the adversary is again $r(M_r, M_b) = \max\{r^{max}(M_r, M_b),\, s\}$.

[23] Which is now the same for both insured and uninsured firms.

expected insurance payment is $q\,M_r$. In order to induce acceptance or given r, the IC constraint of the *firm* that contract terms must satisfy is $r - M_r \leq b - M_b$, and the resulting maximization problem is:

$$\max_{p,M_r,M_b} \{p - qM_r\}, \quad \text{s.t.} \quad r \leq \overbrace{b + M_r - M_b}^{r^{max}(M_r,M_b)} \quad \text{and} \quad U^N \leq U^I(p, M_r, M_b) \quad (3.5)$$

In any equilibrium in which the insurer wants to induce payment of ransom, it also offers full insurance against ransom payments, $M_r(r) = r$. Since M_b is not paid in equilibrium and also does not affect the adversary's ransom offer, the IC constraint is *slack*: any value of M_b can be chosen so that it is satisfied.[24] Thus, for $r \leq b$, the best-response correspondence of the insurer is:

$$BR^I(r) = \begin{cases} M_r = r, \\ M_b \in [0, b], \\ p : U^I(M_r, M_b, p; r) = U^N(r) \end{cases}$$

I emphasize that firms' outside option of not buying insurance yields utility $U^N(r) = (1-q)u(w) + qu(w-r)$, for $r \leq b$, which in the game without observed contracts now depends on the ransom demand r.

Equilibrium.

In equilibrium the insurer's and adversaries' conjectures must be correct, and putting the two best responses together: $M_r(r) = r = BR^A(M_r, M_b) = b + M_r - M_b$ and the second equality must hold in any equilibrium in which ransom is paid. Thus, the only value of M_b consistent with an equilibrium in which ransom is paid is $M_b = b$. Finally, for ransom to be paid in equilibrium, it must be that the adversary earns payoff higher than s, i.e. $M_r \geq s$.

In the Appendix, I prove that if $b > s$, then a continuum of pure-strategy equi-

[24] We verify ex post that the equilibrium value is also positive.

libria with ransom acceptance exists. For every value $M_r \in [s, b]$, there exists an equilibrium in which all firms buy insurance and breached firms accept the ransom offer $r = M_r$. The insurer pays M_r every time a firm is breached. Given the above contract terms, the adversary optimally asks for ransom $r = M_r \geq s$, and $r^{max}(M_r, M_b) = M_r = r$ so that the firm is *indifferent* between acceptance and rejection of the demand. The insurer cannot affect the ransom offer and best responds by offering a contract that provides full insurance, $M_r = r$. The *strategic complementarity* between M_r and r is the source of the equilibrium multiplicity: higher ransom demands increase the amount of M_r required to provide full insurance and higher M_r induces higher ransom demand by increasing insured firms' willingness to pay ransom, $r^{max}(M_r, M_b)$.

In the Appendix, I also show that under $b > s$, there also exist equilibria with *rejection* of ransom, in which the adversary makes "unreasonably" large demands. I offer arguments based on *trembling* and robustness to *counteroffers* to disregard these equilibria and focus on the ones presented in this section.

If $b < s$, there is rejection of ransom in equilibrium and the adversary earns revenue via the sale of data. Full insurance against business interruption is offered in equilibrium, $M_b = b$. As in the corresponding case of Proposition 12, there is again multiplicity of equilibria, since any M_r and r that are consistent with rejection of the ransom offer can be part of an equilibrium. As in the case of observed contracts, all equilibria under $b < s$ are payoff equivalent.

Proposition 15. *In the game with unobserved contracts there exists a pure-strategy SPNE, and firms buy insurance in equilibrium.*

- *If $b > s$: there exists a continuum of equilibria in which the firm accepts the ransom offer. These equilibria can be indexed by the equilibrium value of M_r. In every such equilibrium, the insurer sets $M_b^* = b$, $M_r^* \in [s, b]$, and premium to fully extract firm's surplus, $U^I(p^*, M_r^*, M_b^*) = U^N(r^*)$. Adversaries make*

ransom offers $r^* = M_r^*$. Firms accept the ransom offer. Off-path, if a firm does not purchase insurance and is breached, it accepts the ransom offer, too.

- *If $b < s$: the insurer sets $M_b^* = b$, $M_r^* \in [0, s]$, and premium to fully extract firm's surplus, $U^I(p^*, M_r^*, M_b^*) = U^N(r^*)$. Adversaries make ransom offers $r^* = s$. Firms reject the ransom offer. Off-path, if a firm does not purchase insurance and is breached, it rejects the ransom offer, too.*

Welfare comparison across equilibria. To compare welfare of firms across the multiple equilibria under $b > s$, we can simply compare the off-path expected utility uninsured firms in each equilibrium. The value of the outside option is $U^N(r^*) = q\,u(w - r^*) + (1 - q)u(w) = q\,u(w - M_r^*) + (1 - q)u(w)$, so that the equilibrium in which firm's utility is maximized is the one with $M_r = s$. However, the *opposite* holds for the insurer's expected profit. In the Appendix, I prove the following:

Proposition 16. *For $b > s$: in the game with unobserved contracts, equilibrium expected profit of the monopolist insurer is higher in equilibria with higher M_r. In contrast, equilibrium expected utility of firms is higher in equilibria with lower M_r.*

In light of the above result, we can think of the insurer and adversaries as playing a coordination game; they are both better off when adversaries ask for high ransom and insurance against high ransom levels is provided, but that is at the expense of firms.

Comparison with observed contracts case. Focusing on the case of $b > s$, notice that there exists an equilibrium with $M_r^* = s,\ M_b^* = b$, which is the contract offered in the unique equilibrium under observed contracts (case $b > s$ of Proposition 12). Even though the ransom offer made to the (fully) insured firms is the same across the two equilibria, the *premium* charged by the insurer is different across the two cases. The reason is that the value of the outside option

available to firms is different. Off the equilibrium path, uninsured firms face ransom $r^* = M_r \leq b$ in the equilibrium of Proposition 15, thus their expected utility is *higher* relative to the equilibrium of Proposition 12 in which they face ransom $r^* = b$. Intuitively, off-path, uninsured firms benefit from a positive externality from insured firms' decision to buy insurance, because the adversary's optimal ransom demand is lower for firms that are insured with the equilibrium contract. In the case of observed insurance contracts, there is no such externality because ransom offers are made on the basis of each firm's true contract. Since in any of the aforementioned equilibria firms are indifferent between purchasing insurance and not doing so, firms are *better-off* in any equilibrium with unobserved contracts, relative to the unique equilibrium with observed contracts of Proposition 12. In contrast, adversaries make the maximum possible amount in the equilibrium without insurance, $r = b$. I combine this discussion with the result in Corollary 2 to obtain the following:

Corollary 3. *Under $b > s$: in every equilibrium with unobserved contracts of Proposition 15, firms are better off than both (a) under the unique equilibrium with observed contracts of Proposition 12 and (b) the unique equilibrium without an active insurer, $U^{unobs}(M_r) \geq U^{obs} = U^0$. Adversaries are better off than in the equilibrium of Proposition 12 but worse off than in the equilibrium without an active insurer, $\pi^{obs} \leq \pi^{unobs}(M_r) \leq \pi^0$.*

So even if for $M_r > s$ firms would benefit from moving to an equilibrium with lower M_r, they are nevertheless better off under unobserved contracts relative to the case of observed contracts. Perhaps counterintuitively, adversaries *also* prefer equilibria with unobserved contracts. Of course, holding the insurer's reaction fixed, adversaries would be indifferent between observing insurance contracts and not doing so, since in equilibrium they know the single insurance contract offered. However, if contracts are indeed observed, the insurer finds it profit-maximizing

to use them as a commitment device that depresses equilibrium ransom.

Regulatory Intervention. According to the last result, even if contracts are unobserved, the creation of an insurance market remains welfare-enhancing for firms and welfare-reducing for hackers across all equilibria. However, unlike in the case of observed contracts, there may be grounds for regulating insurance for ransom payments, in particular restricting it so that it remains close to the value that hackers obtain if bargaining fails: this value will mainly depend on (1) the type of data they capture and how much they can profit from exploiting it or selling it and (2) potential direct benefits of causing harm to the firm, which should be particularly relevant for foreign state-backed cyber attacks.

Value of insurance. Insurance contracts in this model offer *risk-sharing* value to firms, as usual: in every equilibrium of Proposition 15, firms are fully insured, regardless of whether they pay ransom or not. But, when $b > s$, insurance also has *commitment* value: adversaries expect the firms to be insured which affects the ransom demand they make, even under unobserved contracts. Since in equilibrium $r = r^{max}(M_r, M_b) = b-(M_b-M_r)$, we can proxy the magnitude of the commitment value by the equilibrium difference of $M_b - M_r$; since that difference is positive in every equilibrium of Proposition 15, firms' willingness to pay ransom decreases and the commitment value is positive, strictly so if $M_r < b$. Under the presence of a fully extracting monopolist, firms in the unobserved contracts model do not appropriate *any* of the risk-sharing value of insurance but they do appropriate the *entire* commitment value. That is precisely because contracts are unobserved, so off-path uninsured firms face the same ransom as insured ones. The greater $M_b - M_r$ is, the lower the equilibrium ransom, and the greater $U^{unobs} = U^N$.[25]

[25] I define the commitment value of insurance as the change in a firm's expected utility because of the change in the equilibrium ransom, ignoring any insurance payments to the firm. In every equilibrium of Proposition 15, the change in a firm's utility relative to U^0 is precisely equal to the commitment value. Since we can again show that all of these equilibria exist under various forms of competition, it is intuitive that the greater the extent of competition, the more of the risk-sharing value of insurance firms will also appropriate in equilibrium.

Remark. The comparison of Propositions 12 and 15 echoes the results of Bolton and Scharfstein [1990], who find that unobservable contracts between a principal and an agent are less effective as pre-commitments than are observable contracts.[26] In their equilibrium with observed contracts, an investor (principal) can effectively use the financing contract it signs with a startup to deter predatory behaviour by an incumbent (who is analogous to the adversary in my context).

But this is not possible in equilibrium when contracts are unobservable, because the principal has a *dominant strategy* when the contracts are unobserved by the incumbent, and the incumbent's optimal response to the dominant strategy is to prey. This is in contrast to my model with unobserved contracts, in which the profit-maximizing insurance contract changes depending on the strategy of the adversary. Ransom demand of the adversary and provision of ransom-insurance strategic complements, leading to the multiplicity result I obtain in Proposition 15.

Random contract observability

A reader may wonder at this point whether the multiplicity of equilibria persists when the adversary can discover the contracts with some exogenous probability $k \in [0, 1]$. If a firm is not insured, the adversary discovers nothing. For interior k, insured firms that are breached face uninformed ransom demand r^u with positive probability $(1-k)$, and with complementary probability informed ransom demand $r^i = \max\{M_r + b - M_b, s\}$. I make the natural assumption that the insurer cannot condition M_r, M_b on whether the adversary has discovered the contract or not. In the Appendix, I show that the multiplicity persists for every $k < 1$. For

[26] In their setting, an investor who chooses a financing contract for a startup faces a trade-off between providing strong incentives to the start-up company and encouraging predatory behavior by an incumbent competitor. The former requires the probability of funding continuation to be dependent on the start-up's first period performance. When the competitor observes contracts, the investor can deter predation by using a financing contract that is less sensitive to performance. In equilibrium, the investor commits to a lower reward for high-performance firms and deters preying.

every $k \in [0,1]$, and for $r(k) := k\,s + (1-k)\,b$, any ransom $r^u = r^i \leq r(k)$ can be supported in equilibrium. Insured firms accept *both* informed or uninformed offers.

Proposition 17. *Assume $b > s$. For any value of k, there exists a multiplicity of equilibria in all of which insured firms accept the ransom offers they face. In equilibrium, $r^u = r^i = M_r^*$, with $M_r^* \in [s, r(k)]$, $M_b^* = b$, and p^* is set to make the insured firms indifferent.*

If $r^u > r(k)$, then the best response of the insurer is to induce *rejection* of the uninformed ransom demand, and no such equilibrium exists under $b > s$. Note that $\lim_{k\to 1} r(k) = s$ which corresponds to the case of observed contracts and $\lim_{k\to 0} r(k) = b$, which corresponds to the case of unobserved contracts. Greater values of k, i.e. greater probability that contracts are observed, monotonically reduce the extent of equilibrium multiplicity, but equilibrium multiplicity arises *for any* $k < 1$.

Remark. For any $k > 0$, $M_b = b$ is a dominant for the insurer: At $k = 0$, even though $M_b = b$ is the sole equilibrium outcome, the insurer is *indifferent* across $M_b \in [0, b]$ in equilibrium. This indifference the indifference is broken toward $M_b = b$ as soon as $k > 0$.

3.6 Liquidity Constraints

An alternative reason why firms demand insurance for ransom payments may be liquidity constraints: firms that face ransom offer r may simply not have enough available funds to pay the adversary, even if they would find it optimal to do so. This is an additional reason why ransomware insurance, and insurance more generally, may be valuable, on top of risk aversion. It is plausible that liquidity constraints matter in the context of ransom payments. Firms must make

funds available within a very short time frame, as well as understand how to make payments according to hackers' technical specifications, and while under pressure to contain the impact of the ongoing ransomware attack. As we will see, the presence of liquidity constraints will *qualitatively* affect how the equilibrium welfare of adversaries and firms compares across equilibria with and without an active insurance market. Additionally, this qualitative difference will depend on whether insurance policies are observed or not by the adversaries.

I assume that every firm has access to liquidity $b > \ell > 0$ and an uninsured firm can only pay ransom r if $\ell \geq r$. On the other hand, if an insurance market exists, an insured firm can pay ransom demand r if $\ell + M_r \geq r$. I assume that liquidity constraints do not impact the firm's ability to pay for insurance. To keep things simple, assume that ℓ is a known scalar, known both to the insurer and the adversary. For brevity, I focus on the case of $b > s$.

3.6.1 Benchmark: without an insurance market

Given the above assumptions, in equilibrium without an insurance provider, firms pay to adversaries as much as their liquidity allows, *if* their liquidity suffices to compensate the adversary for not selling the stolen data. In equilibrium, if $\ell > s$, firms transfer ransom ℓ to adversaries, but if $\ell < s$, no ransom is paid: firms suffer harm b and adversaries sell the data for revenue $s \in (\ell, b)$.

Lemma 20. *Without an active insurance provider, there is a unique equilibrium. If $\ell > s$, firms pay ransom equal to their available liquidity, and payoffs are $\pi^0 = \ell$, and $U^0(\ell) = q\,u(w-\ell) + (1-q)u(w)$. Otherwise, they pay do not pay ransom and suffer harm b. Payoffs are $\pi^0 = s$, and $U^0 = q\,u(w-b) + (1-q)u(w)$.*

Thus, a successful adversary earns $\pi^0 = \max\{s, \ell\}$, increasing in ℓ. The firm obtains expected utility U_0, which increases in ℓ once ℓ exceeds s, and the firm no longer suffers b in equilibrium. But for $\ell \geq s$, U_0 becomes *decreasing* in ℓ, since

the ransom paid is $r = \ell$.

Remark. For any liquidity $\ell < s$, firms would be better off if their liquidity increased to any level $\ell \in [s, b)$ that allows it to pay ransom and avoid the business interruption. This direct, positive impact of the ransom-paying option is a key mechanism in the paper of August et al. [2025].

3.6.2 Observed contracts

For the case of $b > s$, it is easy to verify that the optimal provisions $M_r^* = s$ and $M_b^* = b$ of Proposition 12 are again part of the profit-maximizing insurance policy: the insurer will sell a contract that depresses the ransom demand to the adversary's indifference point and also provides full insurance: the equilibrium ransom demand is $r^* = s < M_r^* + \ell$, so that the liquidity constraint *never binds* for insured firms. In equilibrium, adversaries earn s regardless of whether $s > \ell$.

Lemma 21. *When insurance contracts are observed, and $b > s$, there is a unique equilibrium and in equilibrium, ransom is paid for any ℓ. The insurance contract offered is $M_r^* = s$, $M_b^* = b$ and p^* such that $U^I = U^N$.*

Simple comparison with the previous Lemma reveals that adversaries are strictly worse off in the equilibrium with insurance if $\ell > s$, if firms have sufficient liquidity to pay large ransom in the absence of insurance, and equally well off otherwise. By the same argument of Corollary 2, firms are equally well off under the presence of a fully-extracting monopolist relative to the benchmark.

3.6.3 Unobserved contracts

I look for equilibria of the game with unobserved contracts in *pure strategies*, in which the firm pays ransom. The same arguments I appeal to in the proof of Proposition 15 apply again, and I select only equilibria in which $r \leq b$, i.e. equi-

libria in which ransom is paid in equilibrium. The logic that leads to multiplicity in Proposition 15 applies again, and in the Appendix, I show the following:

Proposition 18. *If $b > s$, there exists a continuum of pure-strategy equilibria in which the firm accepts the ransom demand made by adversaries. For every value of $M_r^* \in [s, b]$, there exists one such equilibrium and in every such equilibrium, $M_b^* = b$ and $r^* = M_r^*$. In every equilibrium, the insurance premium is set to make firms indifferent between buying insurance and not doing so. Off the equilibrium path, uninsured firms accept the ransom offer if and only if $\ell \geq r^*$.*

Welfare comparisons. Same as in the case of observed contracts, there can be no equilibrium in which the liquidity constraint binds for insured firms, because the monopolist finds it optimal to offer full insurance. For that reason, the value of liquidity ℓ does not affect which values of ransom-insurance M_r^* and $r^* = M_r^*$ can arise in equilibrium.

However, the value of ℓ affects the off-path utility of firms if they are uninsured, and thus affects firms' expected utility U^{unobs} (via the equilibrium premium p^*) at the particular equilibrium (M_r^*, r^*) being played. Fixing the value of r^*, the welfare of uninsured firms jumps discontinuously as their liquidity reaches $\ell = r^*$ and remains constant on either side of r^*. Additionally, the value of liquidity affects the firm's and adversary's payoffs in the absence of an insurance market, hence affects how the presence of an insurance market changes their payoffs. Whether an adversary is better off or not depends on how the equilibrium payoff compares to that of the equilibrium without insurance supply, $\pi^0 = \max\{\ell, s\}$, and the answer will depend on *which* of the above equilibria is being played.

Corollary 4. *For the case of $b > s$.*

- *If $\ell < s$, $\pi^0 = s$ and an adversary becomes better off in* ***every*** *equilibrium of Proposition 18 relative to the equilibrium without an insurance market.*

- *If $\ell > s$, $\pi^0 = \ell$ and the adversary becomes (weakly) better off in an equilibrium of Proposition 18 if and only if $M_r^* \geq \ell$.*

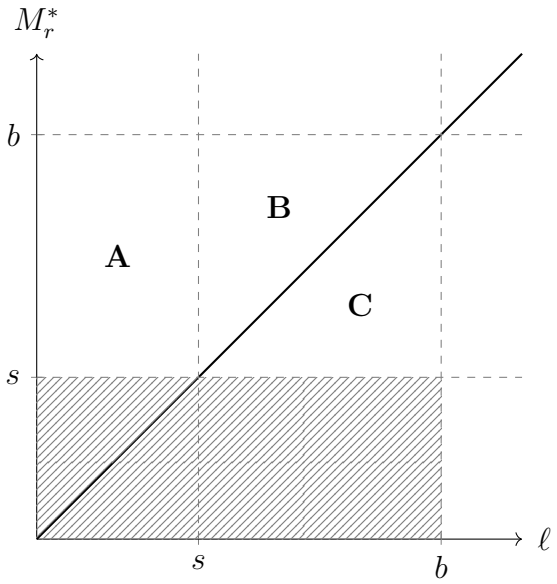
If $\ell < s$, the firm's liquidity constraint is so severe that it cannot even compensate the hacker for the value of their outside option, and there is ransom rejection in the equilibrium without an insurance market, $\pi^0 = s$. Under the presence of an insurer, the revenue of the successful adversary is $r^* \geq s$, so that the adversary becomes (weakly) better off in every equilibrium. If $\ell > s$, then there is acceptance in the equilibrium without an insurer, adversaries are better off only in equilibria in which the amount of insurance for ransomware payments is greater than firms' initial liquidity. Since in such equilibria, $r = M_r$, this implies that adversaries are better off under the presence of an insurer *if and only if* the equilibrium ransom demanded is greater than firms' standalone liquidity ℓ.

Notice that this result comes in stark contrast to the case without liquidity constraints in Proposition 15: in that case, the adversaries are *never* better off under an active insurance market. In fact, this is the first model variant examined with an equilibrium in which hackers are better off relative to the no-insurance profit π^0.

How does the welfare of *firms* compare to U^0, i.e., to the case without an insurance market? With a monopolist insurer who extracts all surplus, firms' expected utility U^* is the same as the off-path expected utility of uninsured firms, U^N. Because insurance policies are unobserved, off-path, firms face the same ransom as on the equilibrium path, $r^* = M_r^* \in [s, b]$ and thus accept if and only if they have sufficient liquidity, i.e. if $\ell > M_r^*$.

Figure 3.1 shows the three possible cases for the comparison of U^N and U^0. In region A, firms are severely liquidity constrained in the absence of insurance, $\ell < s$, and breached firms suffer harm b.[27] With an active insurance market, off path,

[27] With an active insurer, this is the worst possible off-path outcome for uninsured firms, hence



A: Firms equally well off, hackers better off

B: Firms worse off, hackers better off

C: Firms better off, hackers worse off

Figure 3.1: The comparison between the welfare of firms and hackers in an equilibrium with unobserved contracts relative to the equilibrium without insurance depends on firms' liquidity, ℓ and the equilibrium value of insurance for ransom payments, M_r^*.

uninsured firms also cannot pay any ransom in $[s, b]$ and suffer harm b. Hence $U^N(M_r^*) = U^0$ and $U^I(M_r^*) = U^0$, too, for any M_r^*. Firms are equally well off, but hackers are strictly better off in equilibria with $r^* = M_r^* > s$.

In region C, $M_r^* \in [s, \ell]$ and $\ell > s$, and firms are **better off**, i.e. $U^I = U^N > U^0$. The reason is that the equilibrium ransom is *lower* than that of the equilibrium without insurance: the commitment value of insurance contracts, as measured by the difference $(M_b - M_r)$, which acts to reduce equilibrium ransom demands, is sufficiently large relative to compensate for the increase in insured firms' liquidity.

Crucially, the presence of liquidity constraints creates the possibility of a welfare-reducing outcome for firms. This is the case in region B, defined by $M_r^* \in [\ell, b]$ and $\ell \in [s, b)$, and firms are **worse off** in the equilibrium with an insurance provider,[28] $U^I = U^N < U^0$. The outside option U^N becomes lower than U_0 in equilibria with $M_r^* > \ell > s$. When liquidity is not too low, $\ell > s$, breached firms

we immediately see that $U^N \geq U_0$ and firms must be weakly better off.

[28]The result is reminiscent of the main one in Balasubramanian [2021]. In his model, the commitment value of insurance is absent, and by relaxing the liquidity constraint insurance markets *always* lead to higher equilibrium ransom demands.

face ransom ℓ in the absence of an insurance market. In the insurance equilibrium with $M_r^* \in [\ell, b]$, insured firms liquidity constraint is relaxed, and they pay higher ransom than ℓ, thus the ransom demanded by uninsured firms off-path is greater than ℓ, too. Off-path, uninsured firms cannot afford to pay the ransom and suffer business interruption $b > \ell > s$. Hence $U^N < U^0$.

Discussion

In Proposition 15, we already identified equilibria in which M_r^* is too high relative to what minimizes the welfare of adversaries. However, the case of liquidity constraints is the only in which there exist equilibria where adversaries are made better off relative to π^0. This can only occur with unobserved contracts: the insurer does not fully control the commitment value of insurance contracts for the firm in its bargaining with the hacker, since it cannot affect hackers' conjectures, and unlike the case of observed contracts, there exist equilibria with $M_r^* > \ell > s$.

The critical condition such that hackers are better off (and firms weakly or strictly worse off) is that $r^* = M_r^* > \ell$, true in regions A and B. Thus, a message for policy makers is that insurance for ransom payments should not exceed firms' baseline capacity to pay the ransom observed *in equilibrium*. In other words, equilibria in which firms are insured for making ransom payments are welfare-improving, as long as they could also afford to make those same ransom payments without insurance. When this is violated, it is evidence that the equilibrium is in region A or B, and the presence of insurance markets is making adversaries better off. Finally, it holds again that the equilibrium with $M_r^* = r^* = s$ is simultaneously the firm-best and hacker-worst equilibrium. The message of the previous section remains that M_r should be closely tied to the value of stolen data to the hackers.

3.7 Conclusion

In this chapter, I have studied an equilibrium model of the market for ransomware insurance under the presence of strategic ransomware adversaries. The insurance contract specifies different coverage amounts for business interruption and ransom payments made by firms. Purchasing insurance with higher coverage for business interruption provides *commitment value* to the firm in bargaining with the adversary, and insured firms pay *lower* ransom than in the case without insurance markets. In equilibrium, ransom is paid so that firms avoid suffering severe business interruption and coverage for ransom payment is used to provide full insurance to firms.

I study the role of adversaries' ability to observe insurance contracts. With observed contracts, the insurer can directly manipulate ransom demands and in the unique equilibrium, ransom payments become equal to adversaries' outside payoff from selling firms' stolen data. With unobserved insurance contracts, the insurer cannot fully control the commitment value of insurance contracts for the firm, since it cannot affect hackers' conjectures. But even so, the equilibrium insurance contract still provides commitment value for the firm, and this value is greater in equilibria with *lower* coverage for ransom payments. From a policy perspective, these results suggest regulatory intervention that limit coverage for ransom payments can raise welfare and harm adversaries, but does *not* justify banning insurance of ransom payments or the payments themselves. Importantly, in any equilibrium, the welfare impact of insurance markets is always positive for firms and negative for hackers.

This conclusion can change if contracts are unobserved and firms face *liquidity* constraints. Because insurance relaxes those liquidity constraints, there are equilibria in which ransom is higher relative to the pre-insurance equilibrium, and

adversaries are better off. In these equilibria wiht a monopoly insurance market, firms are also made worse off, and the policy implication of the previous section is reinforced: coverage for ransom payments should be limited, and in particular, not exceed firms' standalone liquidity. The latter ensures that firms' welfare is higher in equilibria with an active insurance provider.

Finally, this model can be used to study the interaction between *privacy* and *data-breach* regulations, like the EU GDPR, with firms' incentives in the bargaining subgame. Anecdotally, hackers understand that firms stand to face regulatory scrutiny if their customers' data is leaked and use this as leverage to raise the ransom demanded. This should influence the optimal design of policy concerning both data breaches and ransomware insurance markets.

Appendix 3.A Baseline Model

3.A.1 Proof of Proposition 11

We want to identify the profit-maximizing insurance contract under the constraint that ransom is paid in equilibrium. The relevant IC constraint is $s \leq b + M_r - M_b$. The Lagrangian of the insurer's constrained maximization problem is:

$$\mathcal{L}(p, M_r, M_b) = p - qM_r - \lambda\big(s - \big(b + M_r - M_b\big)\big) - \mu(U^N - U^I(p, M_r, M_b)) \quad (3.6)$$

and I remind the reader that:

$$U^I(p, M_r, M_b) = (1-q)\,u(w-p) + q\,u(w-p-(b-M_b)) \quad (3.7)$$

and

$$U^N = (1-q)\,u(w) + q\,u(w-b) \quad (3.8)$$

First-Order Conditions:

$$
\begin{aligned}
[M_r]: \quad & -q \;+\; \lambda \;+\; \mu \underbrace{\frac{\partial U^I}{\partial M_r}}_{=0} = 0 \implies \lambda = q > 0 \\
[M_b]: \quad & -\lambda \;+\; \mu \frac{\partial U^I}{\partial M_b} \;=\; 0 \implies \mu = \lambda (\frac{\partial U^I}{\partial M_b})^{-1} > 0 \\
[p]: \quad & \mu = -(\frac{\partial U^I}{\partial p})^{-1}
\end{aligned}
$$

where by Proposition 10:

$$U^I(p, M_r, M_b) = (1-q)\, u(w-p) + q\, u(w-p-b+M_b) \tag{3.9}$$

and

$$U^N = (1-q)\, u(w) + q\, u(w-b) \tag{3.10}$$

There is no need to check complementary-slackness conditions, because the first-order conditions reveal that both constraints must bind at an optimal solution. In other words, the adversary must be made indifferent between monetizing via ransom or via the outside option of selling data. The f.o.c. imply:

$$\lambda = q \tag{3.11}$$

$$\mu = \frac{q}{q\, u'(w-p+M_b-b)} \tag{3.12}$$

$$\mu = \frac{1}{(1-q)\, u'(w-p) + q\, u'(w-p+M_b-b)} \tag{3.13}$$

Combining the last two f.o.c. yields:

$$\frac{1}{(1-q)\,u'(w-p)+q\,u'(w-p+M_b-b)} = \frac{1}{u'(w-p+M_b-b)}$$

$$u'(w-p+M_b-b) = (1-q)\,u'(w-p)+q\,u'(w-p+M_b-b)$$

$$u'(w-p+M_b-b) = u'(w-p)$$

$$M_b = b$$

where the last step follows from $u'' < 0$. So, full-insurance against business interruption is profit-maximizing. This already pins down the optimal premium p^* since by Proposition 1, insured firms' utility does not depend on M_r. The optimal premium solves:

$$U^I(p^*, M_r, b) = U^N \iff$$

$$u(w-p^*) = U^N \iff$$

$$p^* = w - u^{-1}(U^N) \iff$$

$$p^* = w - u^{-1}\Big((1-q)\,u(w)+q\,u(w-b)\Big)$$

Given $M_b = b$, the binding IC constraint ($\lambda > 0$) then reveals that $M_r = s$ must hold, which is intuitive: lower values of M_r would violate the IC and induce rejection. Offering a contract with a higher value of M_r would maintain acceptance of the ransom offer in equilibrium of the bargaining subgame, but, as shown above, would not increase the premium a firm is willing to pay. It would, however, increase the expected payment by the insurer, and would thus be less profitable to offer.

3.A.2 Proof of Proposition 12

When the insurer wants to induce no payment of ransom, the constraint becomes $s \geq b + M_r - M_b$, and the Lagrangian of the constrained maximization problem is:

$$\mathcal{L}(p, M_r, M_b) = p - qM_b + \lambda\big(s - \big(b + M_r - M_b\big)\big) - \mu(U^N - U^I(p, M_r, M_b)) \quad (3.14)$$

First-Order Conditions:

$$\begin{aligned}
&[M_r]: \quad \lambda + \mu \underbrace{\frac{\partial U^I}{\partial M_r}}_{=0} = 0 \implies \lambda = 0 \\
&[M_b]: \quad -q - \lambda + \mu \frac{\partial U^I}{\partial M_b} = 0 \implies \mu = q\,(\frac{\partial U^I}{\partial M_b})^{-1} > 0 \\
&[p]: \quad \mu = -(\frac{\partial U^I}{\partial p})^{-1}
\end{aligned}$$

First, the top condition reveals that the relevant IC constraint is now *slack*: since M_r affects neither the utility of the insured firm nor the expected insurance payment, it can be freely adjusted to satisfy the IC constraint for any value of s. Second, combining the last two conditions and repeating the algebraic step from the previous proof implies $M_b^* = b$. Third, and given the above, the optimal premium solves $U^I(M_r, b, p^*) = U^N$. Again, M_r does not directly enter the left-hand side, so p^* is again pinned down by $M_b = b$: This is enough to guarantee full insurance, since the agent's net harm is always $(M_b - b)$, by Proposition 10.

$$p^* = w - u^{-1}\Big((1-q)\,u(w) + q\,u(w-b)\Big) \quad (3.15)$$

When any of the optimal contracts (p^*, $M_b = b$, $M_r \leq s$) is used, the insurer makes expected profit $p^* - qb$.

Appendix 3.B Extensions

3.B.1 Proof of Proposition 13

Once the hacker has successfully breached the firm and seized control of network/-data, the equilibrium ransom is determined by the Nash Bargaining Solution, with β being the bargaining power of the ransomware gang. The insured firm's payoff from successful bargaining is $(-r + M_r)$. Formally:

$$\begin{aligned} r^* &= \arg\max_r \ \big(r - s\big)^{\beta} \big[(M^r - r) - (M^b - b)\big]^{(1-\beta)} \\ &= \arg\max_r \ \big(r - s\big)^{\beta} \big(r^{max}(M_r, M_b) - r\big)^{(1-\beta)} \end{aligned}$$

where $r^{max}(M_r, M_b) = b + M^r - M^b$ is the firm's willingness to pay for ransom, and the first-order condition yields:

$$r^* = (1 - \beta)s + \beta r^{\max} \tag{3.16}$$

Ransom payment occurs in equilibrium if $s \leq r^* \leq r^{max}(M_r, M_b)$, and I look for such an equilibrium. By standard arguments, the profit-maximizing contract must offer **full insurance**, $M_r = r^*$, i.e.:

$$\begin{aligned} &M_r = (1 - \beta)s + \beta(b + M^r - M^b) \iff \\ &\beta(M_b - M_r) = \beta\,(b - s) + s - M_r \end{aligned} \tag{3.17}$$

At the profit-maximizing insurance contract, M_b must be such that the attacker's IC constraint binds, i.e. $r^* = s$. Lowering the equilibrium ransom paid by insured firms raises the premium firms are willing to pay for insurance. At the same time, the insurer's cost is not increased, because M_b is not paid out in equilibrium with ransom payment. The adversary earns his **disagreement payoff** when $r^* = s$

which is equivalent to:

$$(1-\beta)s + \beta(b + M^r - M^b) = s \iff$$
$$\beta(b-s) = \beta(M_b - M^r) \tag{3.18}$$

Combining equations (3.17) and (3.18) shows that the only insurance terms that achieve both full insurance and maximal reduction of ransom demand are: $M_r = s$, $M_b = b$. Thus, these must be part of the profit-maximizing contract. The associated premium is given by

$$p^*(\beta) = w - u^{-1}(U^N(\beta)) \tag{3.19}$$

, where:

$$U^N(\beta) = q\,u(w - h(0,0)) + (1-q)u(w)$$

3.B.2 Proof of Proposition 14

I show that there is a *unique* insurance contract that maximizes expected utility of insured firms, subject to maintaining weakly positive profit for the insurer. Hence, in competitive equilibrium, all insurers will be offering that contract and equilibrium will be in symmetric strategies. Assume that the contract that solves this problem induces payment of ransom in equilibrium. To find the contract that maximizes expected utility subject to the incentive and break-even constraints, I write the relevant Lagrangian:

$$\mathcal{L} = U^I(M_r, M_b, p) - \lambda(s - b - M_r + M_b) - \mu(qM_r - p) \tag{3.20}$$

The associated first-order conditions are:

$$\begin{aligned}
[M_r]: &\quad \frac{\partial U^I}{\partial M_r} + \lambda - \mu\, q = 0, \\
[M_b]: &\quad \frac{\partial U^I}{\partial M_b} - \lambda = 0 \implies \lambda = \frac{\partial U^I}{\partial M_b} > 0 \\
[p]: &\quad \frac{\partial U^I}{\partial p} + \mu = 0 \implies \mu = -\frac{\partial U^I}{\partial p} > 0
\end{aligned}$$

The second condition reveals that as long as higher value of M_b reduces the ransom paid by the focal victim firm without increasing the insurance firm's expected cost of insurance, the IC constraint of the adversary will bind, and $s = b + M_r - M_b$. The third condition implies the break-even constraint will bind and $p^* = q\, M_r^*$. As before, by Proposition 10, the net harm to the firm is always $b - M_b$, and optimality of full insurance implies $M_b^* = b$. The binding IC constraint of the adversary implies $M_r^* = s$. Thus, the equilibrium utility of the (fully) insured firms is $U^I = u(w - p^*) = u(w - q\, s)$.

Finally, we must show that as long as $b > s$, no competitor can profitably undercut by switching to a contract that induces rejection of the offer by an insured firm. I do this by showing that a deviating insurer cannot offer a contract that provides utility greater than $u(w - q\, s)$ to firms while earning weakly positive profits. Even if the break-even constraint binds for the deviating insurer, so that $p' = q\, M_b'$, the highest expected utility he could offer to firms would be lower than $u(w - q\, s)$. With a binding break-even constraint, that expected utility is:

$$q\, u(w - q\, M_b - b + M_b) + (1 - q)\, u(w - q\, M_b)$$

In the event of a successful attack, firms reject the offer, suffer b and are reimbursed M_b. This expected utility is maximized at $M_b = b$ (full insurance). But even at that maximum level, it is equal to $u(w - q\, b) < u(w - q\, s)$, so there is no profitable deviation for the insurer.

Appendix 3.C Unobserved Contracts

3.C.1 Proof of Proposition 15

Best response of insurer Assume the insurer wants to induce acceptance of the offer, in which case expected insurance payment is $q\,M_r$. In order to induce acceptance, the firm's relevant IC constraint that contract terms must satisfy is $r \leq r^{max}(M_r, M_b)$, and the resulting Lagrangian for the insurer's maximization problem is:

$$L(M_r, M_b, p) = p - q\,M_r - \mu(U^N - U^I(M_r, M_b, p)) - \lambda(r - (M_r + b - M_b)) \quad (3.21)$$

where:

$$U^I(M_r, M_b, p; r) = (1-q)u(w-p) + q\,u(w-p-r+M_r) \quad (3.22)$$

$$U^N(r) = (1-q)u(w) + q\,u(w-r) \quad (3.23)$$

and the first-order conditions are:

$$\begin{aligned}
[M_r] \quad & \mu\frac{\partial U^I}{\partial M_r} + \lambda - q = 0 \\
[M_b] \quad & \mu\underbrace{\frac{\partial U^I}{\partial M_b}}_{=0} - \lambda = 0 \implies \lambda = 0 \\
[p] \quad & 1 + \mu\underbrace{\frac{\partial U^I}{\partial p}}_{<0} = 0 \implies \mu = -\Big(\frac{\partial U^I}{\partial p}\Big) > 0
\end{aligned}$$

The condition for M_b implies that $\lambda = 0$, which is intuitive: when M_b is unobserved, the insurer's choice of M_b does not directly affect the ransom offer hence it is costless to change it in order to satisfy the constraint. The last f.o.c. intuitively implies that the second constraint binds and the firm is made indifferent via the

premium. The derivative with respect to the premium is:

$$\frac{\partial U^I}{\partial p} = -(1-q)u'(w-p) - q\,u'(w-p-r+M_r) \tag{3.24}$$

and combining the three f.o.c. yields:

$$qu'(w-p-(r-M_r)) = q\big[(1-q)u'(w-p) + q\,u'(w-p-(r-M_r))\big] \implies$$
$$M_r = r$$

where the last step is implied by the concavity of u. Thus, in any equilibrium in which the insurer wants to induce payment of ransom, it also offers full insurance against ransom payments. Then, the incentive-compatibility constraint of the victim becomes $M_b \leq b$, i.e. insurance against business interruption must be (weakly) incomplete. For the case of $r \leq b$, this will be optimal and instead of inducing ransom rejection which causes greater harm b; in that case, full insurance provision would raise the insurer's cost, and yield lower profit, by the logic of Proposition 12. Thus, for the case of $r < b$, the best-response correspondence of the insurer is:

$$BR^I(r) = \begin{cases} M_r = r, \\ M_b \in [0, b], \\ p : U^I(M_r, M_b, p; r) = U^N(r) \end{cases}$$

In equilibrium the insurer's and adversaries' conjectures must be correct, and putting the two best responses together: $M_r = r(M_r, M_b) = \max\{b + M_r - M_b,\ s\}$. In any equilibrium in which ransom is paid, $\max\{b + M_r - M_b,\ s\} = b + M_r - M_b > s$. The crossing of best responses then implies:

$$M_b^* = b \tag{3.25}$$

Finally, in such an equilibrium, the adversary's IC constraint that must be satisfied to ensure ransom is paid in equilibrium is $M_r \geq s$.

This proves that if $b > s$, then for every value $b \geq M_r \geq s$, the contract $M_r,\ M_b = b,\ p^* = w - u^{-1}(U^N(r))$ and ransom $r^* = M_r^*$ constitute an SPNE strategy profile. In each equilibrium, the insurer pays M_r every time a firm is breached. The adversary best-responds to the insurance contract, the insurance contract maximizes profits, given the ransom demand. Firms are indifferent between purchasing insurance or not and at the bargaining stage are indifferent between accepting and rejecting the ransom offer.

Next, I show that these are the *only* equilibria with acceptance of ransom. There are no equilibria with $M_r = r < s$, because the adversary would not be playing a profit-maximizing strategy. I also prove that there exists no equilibrium in which $r = M_r > b = M_b$. In this case, the insurer would have a profitable deviation. Suppose then that there exists an equilibrium in which $M_b = b$, $M_r = b + \Delta > b$ and all firms buy insurance, hence $r = M_r = b + \Delta$. An SPNE profile must specify sequentially rational strategies off-path, too. Off-path, uninsured firms optimally reject the ransom offer $r > b$, hence the expected utility from not buying insurance is $U^N(r) = qu(w-b) + (1-q)u(w) = U^N(b), \forall r \geq b$. In the candidate equilibrium, the premium is $u(w-p) = U^N(b) \iff p = w - u^{-1}(U^N(b))$ and the expected profit of the insurer is $E\pi = w - u^{-1}(U^N(b)) - q\,M_r < w - u^{-1}(U^N(b)) - q\,b$, for any $\Delta > 0$. If the insurer deviates to $M_r' = 0, M_b' = M_b = b$, the firm optimally rejects the (unchanged) ransom offer because $r^* - M_r' = b + \Delta > b - M_b' = 0$. Since the ransom offer is unchanged, the outside option remains of value $U^N(b)$. The optimal deviation contract also provides full insurance against the relevant risk, hence the optimal deviation premium is again given by $p' = p^* = w - u^{-1}(U^N(b))$. The insurer now pays expected insurance payments $q\,M_b$, thus, the deviation expected profit is $E\pi' = w - u^{-1}(U^N(b)) - q\,M_b > E\pi$, and we have obtained a

profitable deviation.

Second, I prove that if $s > b$, then the equilibrium must feature rejection of the ransom offer. In any equilibrium $r \geq s$. Towards a contradiction, assume that $s > b$ and there is an equilibrium in which **ransom is paid**; then, it must be that the maximum ransom the firm is willing to pay is larger than s, i.e, $r = M_r + b - M_b \geq s > b$. We know that any profit-maximizing contract must feature full insurance, hence $M_r = r$. In equilibrium, the hacker asks for the highest possible ransom that is accepted $r = M_r + b - M_b$, hence the only value of M_b consistent with equilibrium $M_b = b$. This implies that $M_r = r \geq s > b$ must hold in equilibrium.

The insurance premium charged in such an equilibrium must satisfy $U^I(M_r,\, M_b,\, p\,, r) = U^N(r) \iff u(w-p) = U^N(r) \iff p = w - u^{-1}(U^N(r))$ and U^N is *decreasing* in the ransom r. The insurer's expected profit is $p - q\,M_r \leq p - q\,s$ and it is easy to see now that the insurer has a profitable deviation, given the adversary's strategy r, to any contract that induces rejection of the offer r and offers full insurance against business interruption, $M_b' = b$. To induce rejection, the contract must include any $M_r' < r$ and rejection becomes dominant for the insured firms since $r > M_r' + (b - M_b')$. Firms are fully insured again hence are willing to pay up to $p' = p = w - u^{-1}(U^N(r))$. Expected profit is $p - q\,M_b = p - q\,b \geq p - q\,s \iff s > b$, which is true. Hence, any equilibrium under $s > b$ must feature rejection of the ransom offer.

3.C.2 Eliminating additional equilibria

Under $b > s$, there also exist equilibria with ransom *rejection*. The adversary asks for ransom $r^* > b$, the contract offered has $M_b^* = b, M_r^* = s - \Delta > 0$, and the indifference inducing premium is charged, $p^* = w - u^{-1}(U^N(b))$. Nobody has a profitable deviation: insured firms find it optimal to reject since $r > b >$

$M_r + b - M_b = s - \Delta$, and adversaries are indifferent over rejected ransom offers and would earn at most $s - \Delta$ by making a ransom offer that is accepted. The insurer earns $p^* - q\,M_b = p^* - q\,b$ on path and would at most earn $p^* - q\,r < p^* - q\,b$ if it were to offer a contract that induces acceptance of ransom. The counterfactual profit is this because (a) the most profitable candidate deviation is to offer the full-insurance contract that induces acceptance, $M_r' = r^* > b$ and (b) the associated deviation premium remains the same: it is pinned down by the off-path utility of uninsured firms, which remains the same since the ransom they face is unaffected by the monopolist's deviation (unobserved contracts). Hence, the suggested strategy profile is an SPNE.

Even though this equilibrium is **not** Pareto dominated by those of Proposition 15, there are natural reasons to disregard it.

Trembles: First of all, if with small probability ε, firms were to tremble and choose not to buy insurance, $r = b$ would yield strictly greater expected payoff to the adversary for every value of $\varepsilon > 0$.

Counteroffers: Second, in an extended version of the game, uninsured firms off-path have clear incentive to counter-offer to pay $r' = b$, which would make both them and the adversaries better off relative to equilibrium strategies.

For these two reasons, I disregard the pure-strategy equilibrium with the "unreasonably" high ransom demand $r > \max\{M_r + b - M_b,\, b\}$ and restrict attention to those of Proposition 15.

3.C.3 Proof of Proposition 16

In what follows, I evaluate partial derivatives at an equilibrium of the game with unobserved contracts identified in the first branch of Proposition 15. In suhc equilibria, $M_r = r$, $M_b = b$ and p^* is defined as the solution to $v^I(p) - v^N(M_r) = 0$,

which is unique in every equilibrium. I use the ancillary definitions:

$$v^I(p) := U^I(M_r, M_b^*, p) = (1-q)u(w-p) + q\,u(w-p-r^*+M_r) = u(w-p^*)$$
$$v^N(M_r) := U^N(M_r, M_b^*) = (1-q)u(w) + q\,u(w-r^*(M_r, M_b^*))$$
$$r^*(M_r, M_b^*) = b + M_r - M_b^* = M_r$$

Then, expected profit of the monopolist is given by:

$$E\pi = p^* - qM_r \implies \frac{\partial E\pi}{\partial M_r} = \frac{\partial p^*}{\partial M_r} - q \tag{3.26}$$

The implicit function theorem yields:

$$\frac{\partial p^*}{\partial M_r} = -\frac{\partial(v^I - v^N)}{\partial M_r}\big(\frac{\partial v^I}{\partial p}\big)^{-1} = \frac{\partial v^N}{\partial M_r}\big(\frac{\partial v^I}{\partial p}\big)^{-1} > 0 \tag{3.27}$$

where the last equality holds because as M_r is increasing in $[s, b]$ and the equilibrium ransom changes, higher M_r increases equilibrium ransom one-to-one and leaves U^I unchanged, holding the premium fixed. Evaluated at $M_b = M_b^* = b$, these partial derivatives are:

$$\frac{\partial v^I}{\partial p} = -u'(w-p^*) \tag{3.28}$$
$$\frac{\partial v^N}{\partial M_r} = q\,\frac{\partial u(w-r^*)}{\partial M_r} = -q\,u'(w-r^*) \tag{3.29}$$

hence, the slope of equilibrium profit is *positive* if:

$$\frac{q\,u'(w-r^*)}{u'(w-p^*)} - q > 0 \iff$$
$$u'(w-r^*) > u'(w-p^*) \iff$$
$$w - r^* < w - p^* \iff p^* < r^*$$

The equilibrium premium is smaller than r^* in all equilibria in which the firm pays ransom, i.e., for all $M_r \in [s, b]$, otherwise the firms would trivially deviate to not buying insurance. This completes the proof.

3.C.4 Proof of Proposition 17

First, note that Proposition 10 still holds. If the adversary finds out the true contract, the best response is $r(M_r, M_b) = \max\{s, M_r + (b - M_b)\}$ and the insured firm's net payoff is $(M_b - b)$. The adversary will rely on his conjectures with probability $1 - \kappa$, with which he does not find the contract. Fixing those conjectures, an insured firm's payoff is thus:

$$U^I = (1-q)u(w-p) + q\left[\kappa\, u(w-p+M_b-b) + (1-\kappa)\, u(w-p+\max\{M_r-\tilde{r}, M_b-b\})\right]$$

I am operating under the assumption that the insurer cannot condition payments on whether the adversary discovers the contract it or not. With a given choice of M_r, M_b, the insurer determines whether firms pay ransom in *either* contingency. If the hacker is *informed*, insured firms pay ransom if:

$$r(M_r, M_b) - M_r \leq b - M_b \iff r(M_r, M_b) \geq M_r + (b - M_b) \iff M_r + (b - M_b) \geq s$$

I call this constraint IC^i. If the hacker is uninformed, firms pay ransom if:

$$M_r - \tilde{r} > M_b - b \iff M_r + (b - M_b) \geq \tilde{r}$$

I call this constraint IC^u. Both are satisfied when $M_r + (b - M_b) \geq \max\{s, \tilde{r}\} = \tilde{r}$, i.e. if for some contract, firms pay ransom to uninformed hackers, firms also pay to informed hackers, but not vice-versa, i.e. IC^u implies IC^i. In equilibrium, the uninformed hackers also correctly anticipate the insurance contract, so $\tilde{r} = \max\{s, b + M_r - M_b\}$ and the two constraints coincide. This means that in

equilibrium, either firms either pay ransom in both events, or in neither event.

Best response of insurer Suppose the insurer knows the ransom demand by *uninformed* adversaries is $\tilde{r} \in [s, b]$. The ransom demand by informed adversaries is $\max\{b + M_r - M_b, s\}$. By the argument of the last paragraph, the insurer can either use a contract that satisfies both IC^i and IC^u, only IC^i or neither of those.
Case A: Accept offers by both informed and uninformed adversaries.
If the insurer responds with such a contract, the insured firm's payoff is:

$$U^I = (1-q)u(w-p) + q\left[\kappa\, u(w-p+M_b-b) + (1-\kappa)\, u(w-p+M_r-\tilde{r})\right]$$

and the insurer's expected payout is $q\, M_r$. The insurer's Lagrangian is:

$$L = p - qM_r - \lambda\big(\tilde{r} - M_r - b + M_b\big) - \mu(U^N - U^I)$$

The first-order conditions are:

$$[M_r] \quad -q + \lambda + \mu\, q\,(1-k)u'(w-p-\tilde{r}+M_r) = 0$$

$$[M_b] \quad -\lambda + \mu\, q\, k\, u'(w-p-b+M_b) = 0 \implies \lambda > 0$$

$$[p] \quad \mu^{-1} = (1-q)u'(w-p) + q\big[ku'(w-p+M_b-b) + (1-k)u'(w-p+M_r-\tilde{r})\big] > 0$$

For any $k > 0$, i.e., any positive probability with which the hackers discover the contract, the IC constraint for firms facing uninformed adversaries must bind: If that IC is slack, marginally increasing M_b will increase the ex-post utility of insured firms in the event the contract becomes known (and thus their wtp for insurance), and not increase the insurer's payout, since firms are still accepting the ransom. Thus, for any expected $\tilde{r}$, the insurer optimally responds by setting $M_r = \tilde{r} + M_b - b \iff M_r - \tilde{r} = M_b - b$. Combining conditions $[M_r]$ and $[M_b]$ to

eliminate λ yields:

$$\mu^{-1} = (1-k)u'(w-p-\tilde{r}+M_r) + k\,u'(w-p-b+M_b) \iff$$
$$u'(w-p-b+M_b) = u'(w-p-\tilde{r}+M_r)$$

Using $[p]$, we obtain the familiar $M_r = \tilde{r}$. The insurance premium extracted is:

$$p^*(\tilde{r}) = w - u^{-1}(U^N(\tilde{r})) \tag{3.30}$$

and the insurer's profit is:

$$\mathrm{E}\pi^{both} = p^*(\tilde{r}) - q\,M_r = p^*(\tilde{r}) - q\,\tilde{r} \tag{3.31}$$

Case B: Accept offers by informed but not by uninformed adversaries.
The alternative for the insurer is to set a policy (M_r, M_b) such that $s \leq M_r + b - M_b \leq \tilde{r}$, so that the offer of uninformed adversaries is rejected. If the insurer responds with such a contract, the insured firm's payoff is:

$$\begin{aligned} U^I &= (1-q)u(w-p) + q\,[\kappa\,u(w-p+M_b-b) + (1-\kappa)\,u(w-p+M_b-b)] \\ &= (1-q)u(w-p) + qu(w-p+M_b-b) \end{aligned}$$

which is independent of k and M_r because the firm's net loss in both contingencies is $(b-M_b)$ and the insurer's expected payout is $q\,k\,M_r + q\,(1-k)\,M_b$. The insurer's Lagrangian is:

$$L = p - q\,[k\,M_r + (1-k)\,M_b] - \lambda^i\big(s - M_r - b + M_b\big) + \lambda^u\big(\tilde{r} - M_r - b + M_b\big) - \mu(U^N - U^I)$$

The first-order conditions are:

$$[M_r] \quad -q\,k + \lambda^i - \lambda^u = 0 \implies \lambda^i = q\,k + \lambda^u > 0$$

So that IC^i binds and $M_r = s + M_b - b$.

$$[M_b] \quad -q\,(1-k) - \lambda^i + \lambda^u + \mu\, q\, u'(w - p + M_b - b) = 0$$

and combining with the previous condition yields:

$$1 = \mu\, u'(w - p + M_b - b) \tag{3.32}$$

$$[p] \quad \mu = \frac{1}{(1-q)u'(w-p) + qu'(w - p + M_b - b)}$$

and combining with (3.32) yields:

$$M_b = b \implies M_r = s$$

where the value of M_r is implied by the binding IC^i. The insurance premium extracted is the same as in Case A:

$$p^*(\tilde{r}) = w - u^{-1}(U^N(\tilde{r})) \tag{3.33}$$

and the insurer's profit is:

$$\mathrm{E}\pi^{informed} = p^*(\tilde{r}) - q\,[k\,M_r + (1-k)\,M_b] = p^*(\tilde{r}) - q\,[k\,s + (1-k)\,b] \tag{3.34}$$

Compare to the previously derived profit under acceptance of both:

$$\mathrm{E}\pi^{both} = p^*(\tilde{r}) - q\,M_r = p^*(\tilde{r}) - q\,\tilde{r}$$

Putting everything together, the best response of the insurer is always to set $M_b = b$ and $p^* = w - u^{-1}(U^N(\tilde{r}))$. There are two candidate best-response values for M_r: $M_r = \tilde{r}$ if and only if $[k\,s + (1-k)\,b] > \tilde{r}$, to induce acceptance of the uninformed offer; otherwise set $M_r = s$, inducing rejection of the uninformed offer.

Equilibrium

An uninformed demand r^u can thus be part of an equilibrium if and only if $r^u \leq r(k)$, where $r(k) := k\,s + (1-k)\,b$. If $r^u \leq r(k)$, there is an equilibrium in which $M_r = r^u, M_b = b$ and insured firms accept either informed or uninformed offers. If $r^u > r(k)$, then the best response of the insurer is to induce *rejection* of the uninformed ransom demand, and use $M_r = s, M_b = b$. But as argued already, in equilibrium firms either accept both offers or reject both offers. Hence, there is **no equilibrium** with $r^u > r(k)$. Note that $\lim_{k\to 1} r(k) = s$ which corresponds to the case of **observed** contracts and $\lim_{k\to 0} r(k) = b$, which corresponds to the case of **unobserved** contracts. Greater values of k monotonically reduce the extent of equilibrium multiplicity.

Appendix 3.D Liquidity Constraints

3.D.1 Proof of Proposition 18

I focus on equilibria in which the adversary demands ransom $r \in [s, b]$. In equilibria in which the insurance contract induces acceptance of the ransom, the best response of the insurer will always provide full insurance, so that $M_r = r$. Crucially, given this best response, insured firms' liquidity constraint *does not bind*. By assumption, they can use their insurance M_r to cover for the ransom payment.

Just as in the proof of Proposition 15, to see that the candidate strategy profile is indeed an SPNE, notice that (1) adversaries extract ransom equal to firms' willingness to pay, i.e. $r = b + M_r - M_b$, (2) this ransom exceeds adversaries' outside option, $r = M_r \geq s$, and (3) firms are indifferent between accepting or rejecting the ransom demand. Same as in the case without liquidity constraints, the only value of M_b consistent with equilibrium is $M_b = b$. The

The premium is chosen to make firms indifferent between buying insurance and

staying uninsured. The expected utility of uninsured firms depends on the value of ℓ, and on the equilibrium ransom r. Define:

$$h(\ell, r) = \begin{cases} r, \text{ if } r \leq \ell \\ \\ b, \text{ if } r > \ell \end{cases}$$

The off-path expected utility of uninsured firms is $U^N(r; \ell)$ and the equilibrium premium satisfies:

$$u(w - p^*) = U^N(r; \ell) = (1 - q)\, u(w) + q\, u(w - h(\ell, r)) \tag{3.35}$$

Given that the best-response of the insurer is always to provide full insurance, the arguments in Proposition 15 apply verbatim and these are the only equilibria that feature $r \in [s, b]$. There again exist equilibria in which ransom is rejected and $r > b$, but I consider those "unreasonable" and discard them by appealing to the trembling and counteroffer-robustness arguments I discuss in this Appendix.

4 | Prevention and Disclosure of Data Breaches

4.1 Introduction

Cyber attacks and the resulting data breaches are harmful to society, both to firms and to consumers whose data is stolen as a result. Firms need to invest in costly cyber security to avoid data breaches,[1] and when data breaches do occur, society benefits from their *prompt disclosure*. In this paper, I study how regulators should provide firms with incentives to both *invest* in cyber security and also *report* data breaches, especially when these two goals are in conflict.

Prompt disclosure of data breaches allows victims to take *precautions* against the exploitation of their stolen personal information. This is true for people whose credit card or other financial information is stolen, or, for example, for those affected by the recent data breach of the UK NHS IT provider. In the latter, "hackers gained access to ... details of how to gain entry to the homes of 890 people receiving care at home...", BBC News [2024]. Additionally, and at least as importantly, *information sharing* about cyber attacks benefits other firms, who become informed of the state-of-the-art criminal practices, [Kashyap and Wetherilt, 2019], [SentinelOne, 2025]. Specialized governmental entities facilitate such information sharing in both the US and UK.[2]

[1] Two theoretical reasons why we should expect sub-optimally low investment in cyber-security are moral hazard and externalities associated with data breaches. Such externalities may concern personal consumer information or business interruption along a supply chain. Crosignani et al. [2023] find that the propagation of the NotPetya data breach caused a four-fold increase in damages.

[2] The US Cybersecurity & Infrastructure Security Agency is such an entity that facilitates information sharing between private corporations and governmental bodies, see Cybersecurity and Infrastructure Security Agency. In the UK, the Information Commissioner's Office claims that "Every successful cyber-attack that is kept quiet, with no investigation or information

Unfortunately, empirical and anecdotal evidence indicates that data breaches are often reported late — if at all, [Rawson et al., 2023], [Amir et al., 2018].[3] This reluctance to disclose comes at little surprise given the perceived reputational harm for firms that disclose data breaches, see the evidence by Kamiya et al. [2021a]. Prominent examples include the Uber data breach of 2016, which the CEO made explicit attempts to conceal [Gerken, 2023], and the multiple Marriott data breaches of 2014-2020 that were disclosed to regulators with many months of delay [Federal Trade Commission, 2024].

Recognizing the benefits as well as the insufficiency of private incentives, most major jurisdictions already have data-breach notification laws in place. For example, the UK Information Commissioner's Office (ICO) requires that firms report data breaches within 72 hours of becoming aware.[4]

But even though it seems widely accepted that data-breach notification laws are necessary, there is a lack of consensus on their specific design. As Kesari [2024] notes, these laws vary substantially across US states. Design details differ – for example, in the deadline for mandatory disclosure, the specific disclosure requirements, and the level of punishment for non-compliance.

I develop a model to study precisely the design of such regulation. The key trade-off that regulators face is that high penalties for *voluntary* disclosure will on one hand induce larger ex-ante investments to avert data breaches, but will also incentivize firms to delay disclosure of data-breaches that do take place. The existence of this trade-off seems to be understood amongst market participants. As argued in the Sunday Times Palmer [2025], the apparent *under-enforcement* of penalties for data-breached firms in the UK has had the positive effect of higher

sharing, makes other attacks more likely.", National Cyber Security Centre (NCSC).

[3]See Computer Weekly [2023] and the cited industry survey, according to which almost half of organisations that experience critical cyber incidents do not disclose them.

[4]According to the UK ICO, only 61% of reported data breaches were disclosed within the time frame required by the UK GDPR, see Information Commissioner's Office [2025a].

incentives for disclosure.[5]

In my model, once a firm suffers a breach, it privately observes so and chooses how long to wait before disclosing to the regulator. I model this stage as an *optimal stopping* problem. While not disclosing, the firm can potentially *conceal* the breach, but whether that is possible or not depends on the nature of the breach, which the firm is initially uncertain about. At the same time, the firm may be independently exposed, for instance through employee leaks or because consumers are becoming aware of the consequences of the breach. The longer the firm tries to conceal the breach and fails to do so, the more *pessimistic* it becomes about the possibility that it can do so, and after some time of trying unsuccessfully, it voluntarily discloses the breach to the regulator.

The *regulator*, who prefers earlier disclosure and greater investment, plays first and commits to a policy. She chooses two terminal payoffs for the firm: one following voluntarily disclosed breaches and one following exposed ones. An increase in the former payoff will induce earlier disclosure but make it *less* harmful for the firm to suffer a breach, reducing investment incentives: this is the fundamental trade-off the regulator faces. If the regulator wants to induce investment, she should optimally offer firms the largest voluntarily disclosure payoff such that ex-ante investment incentives are maintained. Exposed firms should always receive their limited liability payoff. If this policy still induces large *disclosure delay*, the regulator gives up on providing investment incentives and chooses a lenient policy to induce immediate disclosure instead.

The comparative statics with respect to the probability with which breaches can be concealed are of ambiguous direction. If the firm is initially disclosing with great delay in equilibrium, and concealed breaches impose less social harm than

[5]The author warns regulators who are drafting the UK Cyber Security and Resilience Bill of 2025 "Having to manage a regulatory investigation...closes down clear lines of communications for fear of subsequent enforcement.".

active ones, higher concealability may raise welfare, even if it further reduces disclosure incentives.

I extend the baseline model to capture empirically relevant features of data breaches. First, I show how the regulator's policy changes when firms may discover breaches *after* they have occurred. Such lags in awareness restrict the firm's reaction to a breach, raising the private cost of a breach. This alleviates the trade-off for the regulator and increases the regulator's relative benefit from choosing a policy that induces investment. Second, I consider cases in which breaches that cannot be concealed are also exposed at a higher rate. I show how this can *qualitatively* change the regulator's optimal policy and potentially alleviate the trade-off altogether.

Finally, I study the equilibrium of this model when the regulator can ex-post verify the **delay** with which breaches are disclosed, i.e. the time that the firm has spent trying to conceal the breach. When the only source of private information is the time at which the breach occurs, a *deadline* policy achieves the first-best. The policy provides firms with the lowest possible disclosure payoff *unless* they disclose immediately; thus, the breached firm is made indifferent between disclosing immediately and never disclosing.

If, however, at the onset of a breach, firms receive private signals about concealability, the regulator needs to use disclosure-delay-dependent policies to *screen* the firms with different signals. Under this type of private information, I show that a single-crossing property holds and characterize the set of delay-dependent payoff schedules that the regulator optimally chooses from. Under every such schedule, pessimistic[6] firms disclose immediately, but optimistic ones do so with positive delay. Hastening the disclosure of optimistic firms requires lowering investment incentives.

[6]This is a firm that at the onset of the breach receives a private signal that the breach is less likely to be concealable.

Beyond data breaches, the model should apply well to principal-agent settings with the following key characteristics: (a) firms' unobserved investment determines the state of the world (moral hazard), (b) firms become privately informed about changes in the state, (c) they choose when to disclose this information, and the *timing* of disclosure matters for the principal's welfare, and finally (d) while the firm is avoiding disclosure, it is also collecting information about features of the setting that determine the incentive to disclose. This *learning* component is a main modelling innovation relative to previous literature on the trade-off between ex-ante investment incentives and ex-post disclosure, e.g., Inderst and Mueller [2010] and Levitt and Snyder [1997].

In the very important context of environmental regulation, earlier disclosure can help authorities allocate resources on containing the damage from events like oil leaks or other forms of contamination.[7] In fact, as mentioned in Kim [2015], using lower fines for voluntarily disclosing firms is a known incentive provision instrument in the context of environmental pollution regulation.

4.2 Related Literature

Firstly, my paper contributes to the principal-agent literature that studies the trade-off between provision of ex-ante incentives and ex-post transmission of information. Kaplow and Shavell [1994] show that giving criminals the option of self-reporting can lower investigation costs and achieve the same deterrence rate as without self-reporting. Levitt and Snyder [1997] is the foundational work that studies this trade-off in a principal-agent problem. The agent exerts unobserved effort, then observes a signal of project returns that it may report to the principal. Upon receiving a low report, the principal can choose to cancel the project,

[7] For an example in the UK, there is a clear lack of disclosure by water companies when rivers become polluted with sewage as a result of firm's insufficient ex-ante precautions; Bullough [2022] reports an example of independent "exposure" of the problem.

but incentivizing the bad-news transmission requires weakening the relationship between effort and reward. The trade-off in Inderst and Mueller [2010] is similar: a board chooses the compensation and replacement policies for a CEO who privately learns their match value during his tenure. This makes steep incentive pay optimal so that only good-signal CEOs continue. Severance pay is only used when the CEO's outside value is so low that even under steep incentive contracts, low signal CEOs still want to continue. In Auriol et al. [2023], corporate leniency programs reduce incentives for managers to monitor and report misbehaviour of their employees. They show how the optimal leniency policy is shaped by regulatory capture and, in light of this, discuss existing regulation in different countries. Also conceptually related is the model of Hauser [2023], in which a firm exerts unobserved effort to maintain high quality and can *censor* bad quality signals at a cost. The author finds that for intermediate cost of censorship, there are equilibria in which consumers are better off than in the case without the option of censorship.

Secondly, my work contributes to recent literature that studies the design of *dynamic* self-reporting and inspection policies.

The work of Achim and Knoepfle [2024] is closely related to mine. A firm invests in maintaining compliance and reports on compliance at each time instant. A principal, who *lacks* commitment power, decides when to inspect and how much to penalize firms. In the planner-optimal equilibrium, the inspection policy is *deterministic*. Under a deterministic policy, potential deviations by the principal become immediately apparent to the firm and are maximally costly. Hence, investment and truthful reporting are induced with the lowest inspection cost. In the case of commitment power, the optimal inspection policy is *random*.

The works of Kim [2015], Wang et al. [2016], and Kapon [2022] abstract from moral hazard considerations and feature regulators with commitment power. In all three

papers, the dynamic policies depend on *calendar time* rather than the disclosure delay, which is the case in my model. Kim [2015] asks whether a regulator should conduct inspections at some constant hazard rate, or at known deterministic times. Wang et al. [2016] studies the design of inspection policies and self-reporting rewards and (in contrast to Achim and Knoepfle [2024]) finds that deterministic inspections are optimal. In Kapon [2022], a regulator commits to a dynamic amnesty program to incentivize a stream of criminals to self-report. The criminals have private and stochastically evolving returns from their crime. The paper shows that the optimal policy is cyclical, with amnesty becoming increasingly generous over time before resetting. This structure induces immediate reporting from low-return criminals, while high-return criminals wait to report at the most generous point at the end of each cycle. Related to this literature and to the screening problem I study in Section 4.7, in Halac et al. [2016], a principal hires a firm to explore the viability of a project. The principal wants pessimistic firms to stop exploring early and optimally screens firms with privately known learning rates using dynamic bonus contracts: the firm only gets paid once the project succeeds.

Finally, and immediately relevant to the main application of my model, there is a growing body of empirical work that studies the disclosure of data breaches, e.g., Amir et al. [2018], Kesari [2024], Rawson et al. [2023], Kamiya et al. [2021a]. The work of Amir et al. [2018] interprets the impact of data-breach disclosures on firms' stock value through the Dye [1985] disclosure model. The authors empirically find that data breaches are disclosed with significant delays, and only when the market's perceived probability that a breach has occurred is already sufficiently high.[8] Kesari [2024] studies data-breach notification laws in the US and exploits their staggered adoption across states to show that they reduce instances of identity theft. He also studies the variation in the design of laws across

[8]The available data do not allow the author to distinguish between truly voluntary disclosure versus disclosure of breaches that the market is already aware of – that would be equivalent to "exposure" in my model.

states to identify the most effective provisions. Kamiya et al. [2021a] provides important motivation for my model, as it shows that firms have strong incentives to avoid the disclosure of data breaches: using a data set of self-reported breaches, they find that disclosure does reduce firms' stock value, and much more than what can be explained by penalties, suggesting that reputational concerns strongly disincentivize disclosure in this context.

4.3 Model

Before a breach

A firm (agent) and regulator (principal) are infinitely lived and time is continuous, $t \in [0, \infty)$. While the firm operates, the state is either "safe" or "breached", denoted s and b, respectively. Once, at $t = 0$, the firm decides whether or not to invest in cyber security, $x \in \{0, 1\}$. Positive security costs $C > 0$. The initial state is s and from that, it can either remain s or switch to b. At each point in time while the state is s, the firm earns flow payoff π_s. If the state is s the probability with which the state transitions to b over the time interval $[t, t + dt)$ is $h_x dt$ up to a first-order approximation, with $h_x \in \{h_1, h_0\}$, and with complementary probability the state remains s.[9] Throughout, the principal is assumed unable to observe the investment level and the state of the world. Investment helps deter breaches, so that $h_1 < h_0$.

This part of the model is kept intentionally simple in order to focus on the firm's disclosure decision, once a breach occurs. Next, we proceed with presenting that part of the model.

[9]The arrival time of a breach follows the exponential distribution with hazard rate h_x. The probability of still being in state s at time t is $\int_t^\infty h_x e^{-h_x u} du$.

During a breach

Once a breach occurs, a firm no longer exerts security effort and the flow payoff becomes $\pi_b \leq \pi_s$ presumably because firm's operations become less efficient while breached. While in state b, the firm does not face the risk of an additional breach arriving and there is no additional security choice made at the beginning of state b. From this point onward, time t will refer to **time since the onset** of a breach and not on the calendar time from the previous section (so "time" and "delay" may be used interchangeably).

At every point in time, the firm decides whether to *disclose* the breach to the regulator or *hide*. If it chooses to disclose, it receives instantaneous payoff $D \geq 0$, independent of the disclosure delay, and the game ends. The firm initially lacks information about the type θ of the breach, which can be either "concealable" or not; the initial belief that the breach can be concealed is p_0. If a breach can be concealed ($\theta = 1$) and the firm hides, a "breakthrough"arrives with rate λ, in which case the firm receives delay-invariant, instantaneous payoff of S and the game ends. My primary interpretation of the breakthrough is that the firm successfully destroys evidence of the breach, or destroys evidence that points to its culpability.[10] [11].

While hiding, the firm exposes itself to the risk of being exposed: either by a random audit of the regulator, or by media investigations or simply by consumers independently finding out about the breach, and exposure arrives at rate μ, independently of the type θ and the cumulative hiding time. If exogenous exposure takes place, the firm receives payoff $E \geq 0$ and the game ends. A breach can thus

[10] It can also be interpreted as the firm realizing that the data that has been leaked through the breach is not uniquely held by the firm and thus consumers would not associate the breach with the firm even if they were made aware of it.

[11] Note that in the baseline model, the firm does not choose how many resources to allocate in trying to cover up the breach. The results extend the case in which the firm can only conceal the breach if it exerts effort. In such a case, there is no updating of beliefs in the absence of concealing effort.

end in three ways at some time t since the onset: either it is concealed, for which the time-invariant payoff to the firm is S, or with voluntary disclosure, which yields D, or with random arrival of exposure, which yields E. Following either mode of exit from state b, the game ends.

If after time dt neither a breakthrough nor an audit arrive, the firm finds itself again having to choose between whether to disclose or not, albeit with an updated belief on the type of breach it faces. For belief p_t, Bayes' Rule implies that the decreased posterior before a breakthrough or exposure arrives will be $dp_t = -\lambda p_t(1-p_t)dt$.[12] Thus, given D and E, the firm is facing an **optimal stopping** problem, with state p_t and exit value D. Figure 4.1 illustrates the firm's decision problem.

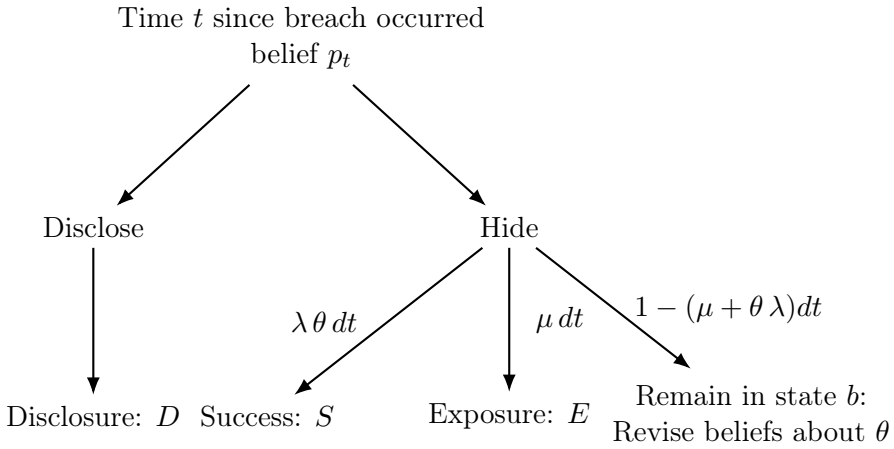


Figure 4.1: The firm's decision tree at time t with belief p_t.

Because the state has a deterministic law of motion, the firm can be thought of as choosing a *deterministic* stopping (i.e., voluntary disclosure) time at the onset of a breach. The state-b expected value, for a firm with prior p_0 that stops at time

[12] This is Bayesian updating with Poisson learning, where breakthroughs arrive at rate λ but only if $\theta = 1$.

T, and given the values D, E, is given by:

$$U_b(T; D, E) := E_{\theta|p_0}\left\{\left[\int_0^T e^{-(\lambda\theta+\mu+r)t}(\pi_b + \lambda\theta S + \mu E)dt\right] + e^{-(\lambda\theta+\mu+r)T}D \;\middle|\; p_0\right\} \tag{4.1}$$

4.3.1 Regulator

At the beginning of the game, and before the firm chooses investment x, the regulator commits to a policy (D, E), i.e. a policy is a pair of termination payoffs following voluntary and involuntary disclosure of a breach. Giving the firm specific values of D or E might require either fining or subsidizing disclosure/exposure, depending on the termination payoffs the firm would earn in the absence of regulation, the "baseline" payoffs D_0 and E_0. I will not be referring to penalties or subsidies, and not think about the baseline payoffs D_0, E_0, but simply refer to D and E as *payoffs*. In my baseline specification I will focus on *simple policies* such that (D, E) are scalars, and in Section 4.7 I will consider *delay-dependent* policies.

Crucially, I assume throughout the paper that the firm is protected by *limited liability*, which means that payoffs must satisfy $D, E \geq 0$. The regulator chooses $D, E \geq 0$ to maximize *welfare*, which I describe next.

4.3.2 Welfare objective

The regulator's expected discounted payoff, which I will refer to as *welfare*, is a function of the firm's choices of investment and disclosure delay. Importantly, I assume that welfare does not depend directly on the precise policy (D, E) chosen. Policy matters only to the extent it influences firm behavior in equilibrium; in other words, any subsidies or penalties required are transfers that do not affect

welfare.[13] I define the function $W_b(T)$ as the regulator's continuation value for transitioning into state b, when she anticipates that the firm will use the stopping time T. Assuming that the regulator uses the same time discount rate as the firm, and for firm investment x, the regulator chooses the policy to maximize ex-ante expected social welfare:

$$W(x,T) = -xC + \frac{w_s}{r+h_x} + \frac{h_x}{r+h_x}W_b(T) \tag{4.2}$$

Flow welfare in state s is w_s and the regulator also internalizes the cost of cyber security investment.

The expression for W_b is similar to that of the firm's continuation value V_b. Remember that a breach can stop in three different ways (concealment, voluntary disclosure, exposure). Keeping things general, the regulator has a different termination payoff for each of the three, namely S^w, D^w, E^w. I discuss the interpretation of these terms and their relation in detail in the following sections.

$$W_b(T) = -L_0 + E_{\theta|p_0}\left\{\left[\int_0^T e^{-(\theta\lambda+\mu+r)t}(w_b+\lambda\theta S^w + \mu E^w)dt\right] + e^{-(\theta\lambda+\mu+r)T}D^w\right\} \tag{4.3}$$

I allow for the regulator to suffer an *instantaneous* social cost from a breach, L_0; it is an *externality* that is completely absent from the firm's b-state calculations. The flow welfare of state b is w_b.

[13] In some models of regulation, as discussed in the book of Laffont and Tirole [1993], there is an additional shadow cost of funds raised via distortionary taxation, thus transfers matter for welfare. I do not make this assumption in order to focus on the investment-disclosure tension. Relatedly, I also assume away costs of imposing fines, which are discussed in Polinsky and Shavell [2000].

4.4 Analysis

I proceed to solve the model with backwards induction. I start by deriving the firm's optimal strategy given a policy (D, E) set by the regulator at the beginning of the game.

4.4.1 Best response of the firm

Given a pair of continuation payoffs (D, E), I derive the firm's strategy which consists of (a) an investment decision in state s and (b) a disclosure threshold $p(D, E)$ (or equivalently, a disclosure delay $T(D, E)$). I first find the firm's optimal policy in state b.

In state b the firm faces a dynamic programming problem, with the state being the current belief $p_t = P_t(\theta = 1)$. As already argued, given that the law of motion for beliefs, conditional on remaining in state b, is *deterministic*, any policy that dictates stopping at some (potentially multiple) level of belief p' induces a deterministic stopping time.[14] Thus the firm is equivalently choosing a stopping (i.e., voluntary disclosure) time T. Define:

$$\gamma(E) := \frac{\pi_b + \mu E}{\mu + r} \quad , \quad \alpha(E) := \frac{\lambda S + \pi_b + \mu E}{\lambda + \mu + r} \tag{4.4}$$

The first term is the expected discounted net present value of playing $T = \infty$, when $p_0 = 0$, i.e. S is never earned, and $\gamma(E) = U_b(\infty; D, E)$ under $p_0 = 0$. Similarly, $\alpha(E) = U_b(\infty; D, E)$ if $p_0 = 1$, and intuitively, $\gamma(E) > \alpha(E) \iff \gamma(E) > S$: that is, when the firm never discloses voluntarily, i.e., $T = \infty$, it prefers $\theta = 1$ if and only if remaining in state b forever is better than the termination payoff S

[14] In the Appendix, I also derive the optimal stopping rule and the analytical form of the continuation value function for state b using the Hamilton-Jacobi-Bellman equation, as in Keller et al. [2005].

that is earned upon a breakthrough.[15]

The firm's *best-response*, when it faces continuation values for disclosure, D, and exposure, E, is to disclose when the posterior drops to $p(D, E)$:

$$p(D, E) = \frac{(\mu + r)(D - \gamma(E))}{\lambda(S - D)} \tag{4.5}$$

If $p(D, E) > p_0$, the firm discloses immediately. We obtain the following characterization of the optimal stopping time $T(D, E)$ as a function of the regulator's policy. In the statement, I make use of the odds ratio $\omega(p) := (1 - p)/p$.

Proposition 19. *For given values of D, E, the following characterizes $U_b(T)$ and the optimal disclosure time:*

1. *If $D > \max\{\gamma(E), S\}$, then $U_b'(T; D, E) < 0$ for all T and $T(D, E) = 0$.*
2. *If $D < \min\{\gamma(E), S\}$, then $U_b'(T; D, E) > 0$ for all T, so $T(D, E) = \infty$.*
3. *If $S > D > \gamma(E)$, then $U_b(T)$ is quasi-concave. If $p(D, E) \geq p_0$, then $T(D, E) = 0$, otherwise $T(D, E) = (1/\lambda) log\Big(\omega(p(D, E))/\omega(p_0)\Big)$.*

Lemma 22. *The disclosure threshold $p(D, E)$ is decreasing in E and increasing in D, so $T(D, E)$ is **increasing** in E and **decreasing** in D.*

The second part of the result is intuitive: greater termination payoff following exposure implies that the firm has greater incentive to hide, i.e. that $p(D, E)$ is lower, and the firm explores for more pessimistic beliefs. On the other hand, a payoff for voluntary disclosure (stopping) payoff implies that the firm discloses earlier.

The optimal stopping policy induces a state-b continuation value function for

[15] These are easily seen if we write:

$$U_b(T; D, E) = p_0\big[\alpha(E) + e^{-(\lambda+\mu+r)T}(D - \alpha(E))\big] + (1 - p_0)\big[\gamma(E) + e^{-(\mu+r)T}(D - \gamma(E))\big]$$

the firm, $V_b(D,E) = \max_T U_b(T;D,E)$. The value of V_b will determine ex-ante incentives to invest in delaying transition to state b. First observe that for any disclosure threshold, the continuation value that the firm expects at the beginning of state b is increasing in both D and E, hence a simple application of the Envelope Theorem reveals the following:

Lemma 23. *The continuation value function $V_b(D,E)$ is increasing in D and E. If $D < \gamma(E)$, then $T(D,E) = \infty$ and changes in D that maintain this inequality leave $V_b(D,E)$ unchanged, and do not affect the firm's optimal strategy. If $D \geq \gamma(E)$, $V_b(D,E)$ is strictly increasing in D.*

Finally, given the continuation value $V_b(D,E)$, the firm chooses $x(D,E) = 1$ if and only if V_b is sufficiently low, relative to the value of remaining in state s. The firm will choose $x = 1$ if:

$$-C + \int_0^\infty e^{-rt}e^{-h_1 t}(\pi_s + h_1 V_b(D,E))dt > \int_0^\infty e^{-rt}e^{-h_0 t}(\pi_s + h_0 V_b(D,E))dt \iff$$
$$V_b(D,E) < \frac{1}{r}\left[\pi_s - \frac{C}{h_0 - h_1}(r+h_1)(r+h_0)\right] := V^{max} \tag{4.6}$$

Define the investment *incentive-compatibility* constraint, $IC_x : V_b \leq V^{max}(C, \pi_s, r, h_0, h_1)$, where V^{max} is defined above and depends on flow profits, the cost of security, and the efficiency of security in avoiding breaches. This constraint implies that in order to maintain investment incentives, the firm must find it sufficiently harmful to suffer a breach (enter state b).

Lemma 24. *Given a pair of continuation payoffs (D,E), the firm chooses $x = 1$ if and only if $V_b(D,E) \leq V^{max}(C, \pi_s, r, h_0, h_1)$*

Trade-off: The source of the trade-off between maintaining effort incentives, i.e. respecting IC_x and incentivizing early disclosure can now be clearly seen. The firm hides until it becomes pessimistic enough about the type of breach it is facing, at

which point it discloses the breach to the regulator to receive D. Holding fixed E, an increase in D has two effects: (a) the optimal stopping time *decreases*, since the prize of disclosure is larger and (b) the b-state expected utility *increases*, which means that the regulator will face a trade-off between decreasing the stopping time and maintaining effort incentives.

4.4.2 Best response of the regulator

Once a breach has occurred, the firm and regulator have potentially different preferred disclosure times, for a given policy in place. In analogy to the above definitions for the firm, we can define γ^w (respectively, α^w) as net present welfare when $\theta = 0$ ($\theta = 1$) and the firm never discloses.

$$\gamma^w := \frac{w_b + \mu E^w}{\mu + r} \quad , \quad \alpha^w := \frac{\lambda S^w + w_b + \mu E^w}{\lambda + \mu + r} \tag{4.7}$$

The logic of Proposition 19 applies for the regulator, too. If $D^w > \max\{\gamma^w, S^w\}$ then the regulator's continuation value $W_b(T)$ is *decreasing* in the firm's disclosure time and immediate disclosure maximizes welfare. In that case, the first-best disclosure time is $T^{FB} = 0$, *regardless* of the firm's investment x. What about the regulator-preferred value of investment?

Lemma 25. *Assume $D^w > \max\{S^w, \gamma^w\}$ which implies $T^{FB} = 0$. Then, a necessary and sufficient condition for the regulator to prefer $x = 1$, for any induced disclosure delay T, is:*

$$r(D^w - L_0) < w_s - \hat{h}\, C$$

where $\hat{h} := \frac{h_0 - h_1}{(r+h_1)(r+h_0)}$

Assuming $D^w > \max\{S^w, \gamma^w\}$ implies $W_b(T)$ is decreasing in T, hence the regulator finds $x = 1$ optimal for all T if and only if she finds investment optimal when $T = 0$, i.e. when disclosure is immediate and the harm from a breach is

minimized. For that reason, in the condition of Lemma 25, the left-hand side is the continuation value for the regulator of entering state b given that the firm will disclose immediately. I will be operating under the following assumptions, unless otherwise stated:

Assumption 1 $D^w > \max\{\gamma^w, S^w\}$

Assumption 2 $r(D^w - L_0) < w_s - \hat{h}\, C$

Before moving on, I remark that the regulator's key constraints that give rise to the trade off are that (1) the firm is protected by limited liability and (2) policies $\{D, E\}$ cannot be conditioned on the value of the firm's investment, i.e. investment is not ex-post verifiable.

The problem becomes trivial in the absence of the limited liability constraint. Intuitively, if $\mu > 0$, an infinitely harsh punishment for exposed firms incentivizes immediate disclosure, for any value of D, which can be chosen to satisfy constraint IC_x. Based on current regulation, limited liability is the natural assumption to make. However, there are discussions in various jurisdictions to increase the maximal loss that firms (or their managers) can suffer for not disclosing a data breach. For instance, there are discussions in EU states to make directors personally liable for breaches, and even suspend them as directors [Palmer, 2025].

In Appendix 4.E, I show that the regulator can always achieve the first-best when she can verify the firm's investment choice after disclosure or exposure and the values of D and E can depend on the verified investment level.

4.4.3 Equilibrium

Under Assumptions 1 and 2, Lemmas 22 and 24 together reveal that $E^* = 0$, i.e. the regulator should maximally punish firms whose breaches are exposed. The regulator faces no trade-off when choosing the value of E: by choosing a lower value she achieves both earlier disclosure and relaxes the effort-IC constraint of

the firm. For brevity, I will be skipping E from function arguments from now on. Then, given $E = 0$, I define two values of D. First, I argue that we can restrict the regulator's *minimum* choice of D to be $D^{min} := \gamma$. This is the largest value of D that is consistent with no disclosure, i.e. solves $p(D, 0) = 0$.

$$D^{min} := \gamma$$
$$V_b(D^{min}) = p_0\,\alpha + (1 - p_0)\gamma$$

Lowering D further has no impact on the firm's choice of T and thus no impact on V_b (Lemma 23), since D is never earned. Next, I argue that we can restrict attention to value of $D \leq D^{max}$, where D^{max} is the *lowest* disclosure payoff that induces immediate disclosure. In other words, it is the unique solution to $p^*(D, 0) = p_0$.[16]

$$D^{max} := \frac{\pi_b + p_0\lambda S}{\mu + p_0\lambda + r} \in (\gamma, \alpha)$$
$$V_b(D^{max}) = D^{max}$$

Since $T(D^{max}) = 0$, the induced continuation value of transitioning into state b coincides with D^{max}. Giving the firm value greater than that would not improve disclosure incentives further, but would increase V_b and thus worsen ex-ante investment incentives, by Lemma 24. We can now describe the equilibrium: a policy and the induced investment and disclosure decisions, $\{D^*, E^*, x^*, T^*\}$.

Proposition 20. *Under Assumptions 1 and 2: the optimal value of the exposure payoff is $E^* = 0$. The following applies to the regulator's choice of $D \in [D^{min}, D^{max}]$:*

- *If $V^{max} < V_b(D^{min})$, then no choice of (D, E) can induce positive investment by the firm. $D^* = D^{max}$ and in equilibrium, $T^* = 0$ and $x^* = 0$*

[16] D^{max} is also the *Gittins Index* of this bandit problem's experimentation arm.

- *If* $V^{max} > V_b(D^{max}) = D^{max}$*, then the optimal policy can induce both positive investment and immediate disclosure.* $D^* = D^{max}$ *and in equilibrium* $x^* = 1$ *and* $T^* = 0$*.*
- *If* $V^{max} \in (V_b(D^{min}), V_b(D^{max}))$*, there are two candidate optimal values of* D^**; either (a)* $D^* = D^{max}$*, which induces* $x^* = 0$ *and* $T^* = 0$*, or (b)* D^* *is the unique value that makes* IC_x *bind, and in equilibrium,* $x^* = 1$ *and* $T^* = T(D^*) \in (0, \infty)$*.*

If the investment IC fails to hold even under the strictest policy D^{min}, then the regulator should again choose D^{max} and induce immediate disclosure, since there is no feasible policy that can yield positive investment in equilibrium. In the second and third cases, $x^* = 1$ can be implemented in equilibrium and we can define as T^{min} the *earliest* disclosure time that can be induced in an equilibrium with $x^* = 1$. In the second case of the Proposition, the investment IC constraint is slack under $D = D^{max}$, and the regulator can achieve $T^{min} = 0$. In the most interesting third case identified above, the regulator chooses between two sources of harm: either delaying disclosure by T^{min} or inducing no investment incentives. She will choose D^* over D^{max} iff: $W(1, T^{min}) > W(0, 0)$.

$$W(1, T^{min}) = -C + \frac{w_s}{r + h_1} + \frac{h_1}{r + h_1} W_b(T^{min})$$
$$W(0, 0) = \frac{w_s}{r + h_0} + \frac{h_0}{r + h_0} W_b(0)$$

and the difference $\Delta W := W(1, T^{min}) - W(0, 0)$ can be written as:

$$\Delta W(T^{min}) = \underbrace{\frac{h_0 - h_1}{(h_0 + r)(r + h_1)}\big(w_s - rW_b(0)\big) - C}_{\text{gain from } x=1 \text{ when } T=0} \underbrace{- \frac{h_1}{r + h_1}\Big(W_b(0) - W_b(T^{min})\Big)}_{\text{gain from earlier disclosure at } x=1} \tag{4.8}$$

In the following section, I will explore how the choice of optimal contract depends

on model parameters. In order to do so, we must understand how parameters affect (1) ΔW directly, holding T^{min} fixed and (2) the disclosure time T^{min}. But before going into comparative statics analysis, let us consider what happens when Assumption 1 is violated, so that the regulator-optimal disclosure time is *positive*.

Case of positive T^{FB}. This is the case if the regulator wants the firm to experiment at $p = p_0$, which is equivalent to $\lambda\, p_0(S^w - D^w) > (\mu + r)(D - \gamma^w)$. Intuitively, if $D^w > \gamma^w$, $T^{FB} > 0$ is only possible if $S^w > D^w$, i.e. if the regulator prefers that the firm exits *privately* rather than via voluntary disclosure. This could, for instance, be true if there is some cost associated with communicating the breach to the regulator that is large relative to the benefit from information sharing. Holding the parameters of the *firm's* problem fixed, a higher value of T^{FB} intuitively means that the regulator's effort-disclosure trade-off is less severe. The value of D necessary to induce stopping at $T^{FB} > 0$ is lower and thus the induced V_b also lower.

Lemma 26. *If the earliest stopping consistent with $x = 1$ satisfies $T^{min} > T^{FB}$, then the regulator can achieve first-best disclosure in equilibrium, and also maintain validity of IC_x by using a lower value of D than the initial D^*.*

Table 4.1: Notation: Firm and Regulator Payoffs

Symbol	**Definition**
h_x	Arrival rate of a breach under investment $x \in \{0, 1\}$
$\pi_s,\ \pi_b$	Flow payoff in the safe state s and the breached state b
S	Lump-sum payoff upon concealment
λ, μ, r	Concealment, exposure and discount rates
$\gamma(E)$	Continuation value in b if never disclosing and $\theta = 0$
$\alpha(E)$	Continuation value in b if never disclosing and $\theta = 1$
$w_s,\ w_b$	Planner's flow payoff in s and b (social counterparts to π_s, π_b)

Table 4.1 – continued from previous page

Symbol	Definition
$D^w,\ E^w$	Planner's valuation of transfers at disclosure/exposure
$\alpha^w,\ \gamma^w$	Planner analogues of α, γ
p_0	Prior belief that $\theta = 1$
$p(p_0, T)$	Posterior after time T
$U_b(T; D, E)$	Firm's value given a fixed disclosure time T
$V_b(D, E)$	Firm's state-b value under optimal disclosure
$T^{\min}$	Earliest implementable disclosure time

4.5 Comparative statics

4.5.1 Regulator welfare

In this section, I study how parameter changes affect $W(x, T)$, holding investment x and stopping time T fixed. The "direct" comparative statics of W with respect to most parameters are straightforward. I will be focusing on the effects of the parameters that govern the rates at which the firm exits state b, i.e. p_0, λ, μ. What is going to be key in deriving and interpreting these comparative statics results is how the regulator's termination payoff depends on the way in which an active breach ends.

Exposure rate. In the plausible case that the regulator values $E^w = D^w$, i.e. in the case where the regulator is indifferent to the mode of disclosure, welfare is increasing in the exposure rate. I show the intuitive result that:

Lemma 27. *If $E^w \geq D^w > \max\{\gamma^w, \alpha^w\}$, then the partial derivative of welfare with respect to the exposure rate, μ, is positive for all $T > 0$, and zero at $T = 0$.*

The second inequality is Assumption 1. Under the condition stated in the result,

we obtain $\frac{\partial^2 W_b}{\partial T \partial \mu} > 0$, which combined with $\frac{\partial W_b(T=0)}{\partial \mu} = 0$ yields the result. The regulator finds delayed disclosure less costly when μ is higher.

Probability of breakthrough. Whether the regulator's welfare increases with higher p_0 depends on the firm's planned disclosure time, T. In particular, it depends on whether the regulator would rather the firm conceals the breach or continues in state b with the plan of disclosing after time T.

Lemma 28. *The partial derivative of welfare with respect to the ex-ante probability of $\theta = 1$, p_0 is:*

- *Negative for all $T > 0$, if $D^w > \gamma^w > \alpha^w$.*
- *For $D^w > \alpha^w > \gamma^w$: there always exists a $\hat{T}^p \geq 0$, such that $\frac{\partial W}{\partial p_0} > 0$ if and only if $T > \hat{T}^p$. In the limit as $T \to \infty$, $\lim_{T\to\infty} \frac{\partial W_b}{\partial p_0} = \alpha^w - \gamma^w > 0$.*

If the regulator prefers concealed breaches to active ones ($\alpha^w > \gamma^w$), the impact of p_0 on welfare depends on the firm's voluntary disclosure time, T. If the firm plans to stop very late, then the harm from a decreased probability of reaching the disclosure time T is dominated by the benefit of less time spent with an active breach. The first effect dominates for early disclosure. If concealed breaches are more harmful than active ones in expectation ($\alpha^w < \gamma^w$), an increase in p_0 always harms the regulator. Note that both cases are consistent with Assumption 1, i.e. under either case, $T^{FB} = 0$. For the (direct) impact of the concealment rate λ on welfare, the intuition is similar to that of the previous result:

Lemma 29. *The following are true about the partial derivative of welfare with respect to the concealment rate, λ.*

1. *For $\alpha^w < \gamma^w$: $\frac{\partial W}{\partial \lambda} < 0$.*
2. *For $\alpha^w > \gamma^w$: there always exists a strictly positive $\hat{T}^\lambda > 0$, such that $\frac{\partial W}{\partial \lambda} < 0$ for $T \in (0, \hat{T}^\lambda)$ and $\frac{\partial W}{\partial \lambda} > 0$ for all $T \in (\hat{T}^\lambda, \infty)$.*

How the payoffs D^w, E^w, S^w, γ^w compare to each other reflects the nature of the setting and in particular the interpretation of the firm's breakthrough that leads to payoff S^w. In the next paragraphs, I discuss these different interpretations of the stopping problem and the relevance for different applications.

Discussion: Interpretation of the breakthrough

Concealment: If we interpret the breakthrough as the firm concealing the breach, with the adverse effects continuing to accrue in the background, but not affecting the firm itself, this is best captured by $S^w < \gamma^w \leq D^w$. The regulator would rather the firm stays in state b, where it has not yet concealed the problem and may be still exposed with rate μ. Disclosure remains the preferred mode of ending an active breach.

Private Fix: On the other hand, if at the arrival of the breakthrough the firm privately fixes the breach, the nautral assumption is $S^w > \gamma^w$. Whether $D^w > S^w$ would in this case depend on whether there are benefits of information sharing (sharing details of how the breach occurred with other firms) relative to the potential cost of communicating a breach to the regulator and consumers. Note that if disclosure is socially costly and the firm can privately deal with the issue comprehensively, $S^w > D^w > \gamma^w$ but as we saw in the earlier section, that would contradict $T^{FB} = 0$: the regulator would rather firms to explore private solutions before disclosing, for sufficiently high values of p_0 and λ.

Opportune disclosure: Values of S^w in the interior of (γ^w, D^w) could correspond to a slightly different interpretation of the breakthrough: one in which it corresponds to the "arrival" of an opportune time for the firm to disclose bad news, so as to minimize their impact on firm value. The papers by Rawson et al. [2023] and Foerderer and Schuetz [2022], specifically for the context of data breaches, empirically find that firms do "bundle" bad news with other information and this alleviates the impact on the share price. Managers seek to distract investors and

increase their processing costs by issuing concurrent disclosures about unrelated events. In that case a breakthrough is followed by a "noisy disclosure". Such disclosures likely fail to deliver the full social benefit, hence $S^w < D^w$.

4.5.2 Earliest disclosure T^{min}

Having understood the direct impact of model parameters on the regulator's welfare, I proceed to study the impact of model parameters on T^{min}, is the earliest equilibrium time of voluntary disclosure subject to $x^* = 1$. Both components are required to then ask how parameters affect the regulator's optimal choice of policy. Throughout, I assume that $T^{min} > 0$, i.e. that the associated value of the voluntary disclosure payoff satisfies $V^{max} \in (V_b(D^{min}), V_b(D^{max}))$.

Take for example the b-state flow profit of the firm, π_b. A decrease in that parameter implies that the firm suffers a flow cost while in state b. Total differentiation shows:

$$\frac{dT^{min}}{d\pi_b} = \frac{\partial T^{min}}{\partial D^*}\frac{\partial D^*}{\partial \pi_b} + \frac{\partial T^{min}}{\partial \pi_b} \tag{4.9}$$

There is a direct effect on the stopping time, holding the principal's policy fixed, and an indirect one via the principal's policy, D^*.[17] The direct effects of state-b parameters on T are mostly straightforward: the firm's optimal stopping time is increasing in π_b, S, as well as p_0 and decreasing in μ. As also discussed in Halac et al. [2016], the direct effect of λ on the stopping time is more subtle. An increase in λ increases experimentation incentives because it makes a breakthrough more likely under $\theta = 1$. The optimal stopping threshold in terms of posterior beliefs, p^* is unambiguously decreasing in λ which tends to increase the firm's stopping time. However, there is an additional *negative* effect on the stopping time because higher λ increases the speed of learning. The firm reaches a given posterior belief faster, because a lack of breakthrough is stronger evidence of $\theta = 0$ when λ is

[17] With some abuse of notation, I am using D^* to denote what is not necessarily the equilibrium value of the policy, rather the policy that induces T^{min}, such that $T^{min} = T(D^*)$.

larger.

Lemma 30. *Holding D fixed, the firm's optimal disclosure time, T, is quasi-concave in the technology parameter, λ. Additionally:*

- *For $\lambda \to \infty$, the slope approaches zero.*
- *For $\lambda \to \lambda_0$, the slope is positive, where λ_0 is the highest value of λ that makes the optimal stopping time equal to zero.*

For the indirect effect of parameters on T^{min}, to obtain the sign of the change in D^*, we use the definition of D^* as the value of D that makes the ex-ante incentive compatibility constraint, IC_x, *bind*.

$$V^{max} = V_b(\pi_b, D^*(\pi_b))$$

where I am being explicit about the dependence of the regulator's policy D^* and of the continuation value on π_b. Following a change in π_b, the principal will optimally change D^* to maintain IC_x validity, and we obtain that $\partial D^*/\partial \pi_b < 0$. This is an intuitive result: ceteris paribus, an increase in π_b causes an increase in V_b and forces the principal to be harsher to firms that voluntarily disclose breaches in order to maintain a binding IC_x: hence T^{min} increases both by the direct effect of π_b but also indirectly because the principal reduces D.

Lemma 31. *For parameter regions such that $V^{max} \in (V_b(D^{min}), V_b(D^{max}))$, the disclosure time $T^{min} = T(D^*)$ is:*

1. *Increasing in π_b, p_0, h_1, C*
2. *Decreasing in π_s, h_0, μ.*
3. *Increasing in λ, for low initial values of T^{min}.*

Note that parameters $\{h_1, h_0, \pi_s, C\}$ do not affect state-b payoffs and thus have no direct effect on T^{min}, but affect the value V^{max} and thus the principal's choice

of D^*. A decrease in h_1 means that the firm is more willing to invest hence the regulator can increase D^* and bring disclosure forward without jeopardizing investment incentives. Similar reasoning applies to h_0 and C. The total effect of λ on the earliest disclosure time T^{min} is the sum of a positive indirect effect, via forcing the regulator to decrease D^* to maintain investment incentives, and an ambiguous direct effect. By the logic of Lemma 30 if T^{min} is sufficiently close to zero to begin with, the total effect of the breakthrough rate on T^{min} is unambiguously positive.

4.5.3 Optimal contract

Putting everything together, we can examine how changes in parameters affect the regulator's welfare under each of the two candidate optimal contracts, and thus the regulator's optimal choice. Define the relative welfare from inducing $x^* = 1$ as the difference of the two sides in (4.8):

$$\Delta W(T^{min}) := -C + \frac{1}{(h_0 + r)(r + h_1)} \Big[w_s(h_0 - h_1) + h_1(r + h_0)W_b(T^{min}) - h_0(r + h_1)W_b(0) \Big]$$

Lemma 32. *Under Assumptions 1 and 2, the regulator's incentive to choose the optimal investment-inducing contract, $\Delta W(T^{min})$, is:*

1. *Increasing in the exposure rate μ, if additionally $(D^w - E^w)$ is sufficiently small.*
2. *Decreasing in π_b.*
3. *Increasing in S^w.*
4. *Increasing in h_0 and decreasing in h_1, C.*

Even though the comparative statics result with respect to the exposure rate μ, is conceptually simple, it is of policy relevance. When the exposure rate or the

regulator's audit capacity is greater, the regulator should be more willing to choose contract that induces investment. When μ increases, not only is the harm from later (planned) voluntary disclosure reduced, but also earlier disclosure can be achieved alongside positive investment. This result echoes the similar finding of Auriol et al. [2023].

An increase in π_b is interpreted as the firm appropriating *less* of the harm from an active breach. The lower the private cost to the firm of an active (or unconcealed) breach, the lower the incentive to disclose early (see Lemma 31) and thus the lower welfare $W_b(T^{min})$ is.

Intuitively, the greater the value of S^w, i.e. the greater the social value of the firm's private exit, the lower is the welfare cost of inducing investment as captured by a later disclosure time. As $S^w \to D^w$, i.e. the gain to information disclosure decreases relative to a private "fix", then the regulator has stronger incentive to use the contract that induces $x^* = 1$, since it is becomes less costly to delay disclosure. The opposite is true if when the breakthrough occurs the firm manages to conceal the breach without stopping the accumulation of harms to society, in which case S^w is low and the regulator has stronger incentive to induce immediate disclosure and give up on investment.

An increase in the efficiency of investment, i.e. a *decrease* in h_1 also means that the regulator has stronger incentive to induce positive investment; that is true not only because of the direct effect on ΔW, but also because of Lemma 31.

Finally, I examine the effect of λ, p_0 on $\Delta W(T^{min})$. As previous results revealed, the direction of the direct effects of these parameters on $W_b(T^{min})$ but also of the indirect effects via T^{min} can depend on model parameters and on the initial value of T^{min}. As we know from Lemmas 28 and 29, both direct effects on $W_b(T^{min})$ are *negative* regardless of parameters, for low values of initial T^{min}. Combined with Lemma 31, this implies $W_b(T^{min})$ and $\Delta W_b(T^{min})$ is decreasing in both λ

and p_0 when starting from low values of T^{min}. Increases in either parameter will increase the benefit for the regulator to choose the contract that induces immediate disclosure, and give up on investment incentives.

Lemma 33. *For parameters such that T^{min} is low, the total effect of increases in p_0 and λ on $\Delta W(T^{min})$ is negative.*

4.6 Extensions

4.6.1 Delayed awareness of breaches

In the baseline model, I assume that a firm realizes immediately when a breach occurs. But as Kamiya et al. [2021a] and Kashyap and Wetherilt [2019] point out, and the already mentioned Marriott case study demonstrates, this is not necessarily the case. In this section, I examine the impact on the regulator's optimal policy of such delays in the firm's *awareness* of a data breach. In particular, I assume that a firm either discovers a breach immediately, or with exogenous probability ν does so with positive delay $d > 0$.[18]

A key assumption in this section is that collecting information or attempting to conceal the breach requires active effort by the firm, i.e. cannot be done before the firm has become aware of the breach. In other words, $\lambda = 0$ while the firm is unaware of an active breach, regardless of the type of breach, θ. Hence, the initial belief is p_0, regardless of the delay of awareness. The implication is that equilibrium time from awareness of the breach to voluntary disclosure is *unaffected* by the timing of awareness, because the initial belief is independent of d and the other elements of the stopping problem are independent of the time spent in state b. Finally, I assume that pre-awareness, a breach can be exposed at the usual rate

[18]The firm also plausibly influences the monitoring of its systems and thus the delay with which it realizes that a breach has occurred. I abstract away from endogenous monitoring in this extension.

μ.

When the breach occurs, the regulator obtains expected continuation payoff $\mathbb{E}_d[W_b(T;d)]$, where the expectation is taken over the realization delay, d. The first argument is the time *after* the firm has realized the breach and before it reports it to the regulator. We can write the regulator's continuation utility for state b, when the firm discloses at time T *after awareness*, as:

$$\begin{aligned}\mathbb{E}_d[W_b(T;d)] = & \overbrace{(1-\nu)}^{\text{immediate awareness}} \quad W_b(T;0) + \nu\, W_b(T;d) \\ & = (1-\nu)W_b(T;0) + \nu\Big[\big(1-e^{-(\mu+r)d}\big)\gamma^w + e^{-(\mu+r)d}W_b(T;0)\Big]\end{aligned} \tag{4.10}$$

If the firm never realized the breach occurred, the regulator's continuation payoff would be $W_b(T;\infty) = \gamma^w$, since the firm would *never* be able to voluntarily disclose the breach. From the time the firm realizes the breach onwards, the regulator earns continuation payoff $W_b(T)$, defined same way as before. I want to understand how parameters ν, d affect the regulator's choice of policy. The welfare loss from maintaining investment incentives is that disclosure happens with a delay. But, intuitively, when the expected delay with which the firm realizes the active breach is large, this benefit becomes of lower magnitude. Hence an increase in the expected delay tilts the regulator's decision towards providing investment incentives, holding T^{min} fixed.

Additionally, it is easy to see that the indirect effect is also positive: T^{min} is decreasing in d and ν. Since a delay in the realization restricts the firm's ability to react to a breach, it reduces the firm's continuation value V_b, hence increases the incentive to invest.[19] Both effects thus point towards the same direction, and we can make the following statement about the regulator's incentive to induce positive investment, ΔW.

[19]Because neither variable affects the value of V^{max}, see (4.16).

Proposition 21. *Following a marginal increase in either ν or d:*

- *The disclosure time T^{min} decreases.*
- *The relative social welfare from inducing investment, ΔW increases, holding T^{min} fixed.*
- *Hence, ΔW increases unambiguously when taking both effects into account.*

Of course, this does not mean that the regulator *prefers* larger expected realization delays; holding all else fixed, the regulator may dislike delays in the firm's realization.[20] Nevertheless, we have shown that in environments with greater expected realization delays, the regulator has greater incentive to use the contract that implements both investment and disclosure.

4.6.2 Exposure rate depends on type of breach

Under my primary interpretation of the model, the firm is learning whether it can conceal the breach or not, with $\theta = 1$ indicating a breach that can be concealed. It may be natural for some settings to assume that the exposure rate μ also depends on the underlying value of θ, with $\mu(0) > \mu(1)$, so that breaches that *cannot* be concealed are also breaches that are exposed with *greater* probability.[21] This may be a relevant case for the data-breach context. Imagine, for example, a firm that learns whether stolen data has been exploited in a way that will become publicly known to consumers. Consumers can become aware of stolen passwords being used to grant unauthorized access to online services, but may never know that the particular company suffered a data breach if their data is used for phishing, or identity theft. To simplify expressions without losing any insight, I present how the firm's policy changes under this variation of the stopping problem for $\mu(1) = 0$

[20] This will be true for low exposure rate, μ. If μ and λ are large, and $S^w << D^w$, the regulator may prefer that the firm does not have the opportunity to conceal the breach, and higher d may increase welfare.

[21] The baseline corresponds to $\mu(0) = \mu(1)$.

and use the notation $\mu_0 = \mu(0) > 0$.

When $\mu_0 > \mu_1 = 0$, as the firm hides and neither a breakthrough nor exposure arrives, it may, intuitively, become more *optimistic* about the probability of $\theta = 1$, since the probability of exposure is greater under $\theta = 0$. The law of motion for beliefs is $dp = -(\lambda - \mu_0)p(1-p)dt$, i.e. beliefs drift downwards if and only if $\lambda > \mu_0$. In that case, the firm's optimal policy remains qualitatively the same as in Proposition 19. For that reason, I focus here on the case of $\lambda < \mu_0$. The longer that consumers do not become aware of the breach, the *less* likely it is that they ever will.

If $\mu_0 > \lambda$, the firm's best response is qualitatively different: there exists a threshold $\hat{p} > 0$, such that the firm stops immediately, if $p_0 < \hat{p}$, otherwise *never* discloses, and becomes progressively more optimistic that $\theta = 1$ over time as it remains in state b. This change has stark implications for the regulator's optimal policy. By the same logic as before, $E^* = 0$ remains optimal. But, in contrast to Proposition 20 exists a single policy $\hat{D}$ that is *always optimal*. That is the smallest value of D that induces immediate disclosure. If at $D = \hat{D}$, the IC_x constraint of the firm is satisfied, the regulator achieves the first-best outcome. Otherwise, *no* policy can induce positive investment, even if it sacrifices disclosure incentives. The reason is that under $\hat{D}$, the firm is already indifferent between disclosing immediately or never. Since it can always guarantee itself the payoff from never disclosing, no policy can induce a lower continuation payoff V_b than policy $\hat{D}$.

Proposition 22. *For $\mu_0 > \lambda$ and $\mu_1 = 0$: the regulator's optimal policy is to use $E = 0$, and $D = \hat{D}$, given by:*

$$\hat{D} = p_0 \frac{\pi_b + \lambda S}{r + \lambda} + (1 - p_0) \frac{\pi_b}{\mu + r_0} \tag{4.11}$$

In equilibrium, $T^* = 0$. Investment is positive if constraint IC_x is satisfied.

The optimal policy makes the firm *indifferent* between disclosing immediately or never. In contrast to the baseline, there is a single policy that is always optimal and there is *never* any disclosure delay in equilibrium: the regulator effectively no longer faces a trade off.

4.7 Delay-dependent penalties

After a data breach becomes disclosed or exposed, it may be possible to understand when data was stolen, for instance via IT forensics or by looking at when stolen consumer data was first exploited. For the remainder of the paper, I depart from the baseline to study how a regulator would adapt their optimal policy if they had some ex-post verifiable information on the delay with which a breach is voluntarily disclosed or exposed.

In this section, I assume that upon disclosure or exposure, the regulator can ex-post verify the *delay* with which the firm disclosed the breach and can write enforceable contracts $\mathcal{C} = \{D(t), E(t)\}_{t\geq 0}$, where t is time elapsed since the breach occurred. The limited liability constraint remains, i.e. I focus on $D(t), E(t) \geq 0$, for all $t \geq 0$ and all parameters are commonly known from the beginning of the game. I maintain Assumptions 1 and 2 so that the objective $W(x, T)$ is increasing in x and decreasing in T. This implies that the equilibrium exposure payoff is always the minimum possible, i.e. $E(t) = 0$, for all t.

Given $E(t) = 0$ for all t, the state b expected discounted utility for a firm of prior p_0 that never discloses is $(\alpha - \gamma)\, p_0 + \gamma$, where α and γ are the terms defined in (4.4).[22] No matter the disclosure payoff schedule, $D(t)$, the firm can always attain this state-b payoff by never disclosing, hence $(\alpha-\gamma)\, p_0+\gamma$ is the value of the firm's outside option, the minimum continuation value V_b that the firm can obtain in

[22]This is the value of $V_b(D^{min})$ in the previous section. As argued above, under limited liability the firm can never earn lower V_b than this. Hence the regulator who is restricted to scalar policies is wlog using $D \geq D^{min} = \gamma$.

equilibrium. The following Proposition identifies a contract that simultaneously (1) induces immediate disclosure and (2) gives the firm this minimum continuation value, maximizing ex-ante incentives for investment.

Proposition 23. *[Deadline Contract] Assume that the regulator has objective $W(x,T)$, increasing in x and decreasing in T. Then, if the regulator can choose any delay-dependent contract that respects limited liability, it is optimal to use the following:*

$$D(t) = \begin{cases} (\alpha-\gamma)p_0+\gamma & if\, t=0, \\ 0 & if\, t>0 \end{cases}$$

$$E(t) = 0, \ \textit{for all } t$$

That contract offers disclosure payoff of $D(0) = (\alpha-\gamma)p_0+\gamma$ to any firm which immediately discloses a breach, and disclosure payoff equal to zero for disclosure with any positive delay. Under the above contract, the firm chooses $T^* = 0$ and the firm's continuation value is $V_b = D(0)$.

Proposition 23 points to a contract that is optimal **regardless** of whether the regulator wants to respect IC_x. Consider the two cases: (a) if $V^{max} > (\alpha-\gamma)p_0+\gamma$, then under the above contract IC_x is respected and we have both positive investment and immediate disclosure (b) if $V^{max} < (\alpha-\gamma)p_0+\gamma$, then *no other contract* that can be used will respect IC_x, since even the lowest possible value of V_b is not low enough to incentivize ex-ante investment by the firm. In that case, since the regulator does not directly care about the value of D and E, any contract that yields prompt disclosure is optimal from the principal's perspective. Thus, Proposition 1 tells us that the regulator who prefers prompt disclosure would always incentivize immediate disclosure and faces *no trade-off* between ex ante and ex post incentive provision.

Comparison to Proposition 20. Is it useful for the regulator to condition

penalties on the delay of realization? Does she achieve better outcomes relative to the previous section? First of all, remember from Proposition 20 that if the regulator wants to induce $V_b = (\alpha - \gamma)\, p_0 + \gamma$ using a delay-independent, scalar penalty, she must use policy $D = D^{min} = \gamma$, yielding no disclosure at all, $T = \infty$. Conversely, in order to achieve immediate disclosure with scalar policies, the regulator must use $D = D^{max}$ which can be shown[23] to be greater than D^{min}, and for that reason yields lower ex-ante incentives than the contract of Proposition 23. More generally, whenever in the equilibrium of Proposition 20 the regulator chooses a contract to induce $x^* = 1$ and $T^* > 0$, she is strictly better off by the ability to condition the payments to the delay of disclosure.[24] The new contract will also achieve $x^* = 1$, and maintain $T^* = 0$.

4.7.1 Heterogeneity of breaches

The stark outcome of Proposition 23 is a good benchmark, but it is worth it to use the developed insight to investigate how the policy changes in the presence of *private information*. I extend the model by assuming that the firm, upon getting breached, not only privately finds out about the change in state, but also receives a private signal that determines its belief at $t = 0$ about θ, the type of breach it is facing. The *time-zero belief* is $p_0 \in \{L, H\}$, with $L < H$ and $P(p_0 = L) = q$. The regulator, instead, only knows the ex-ante distribution over time-zero beliefs, which is also what is known to the firm before the arrival of a breach. The interpretation of this is simple: upon suffering a breach, the firm has an informational advantage over the regulator in terms of assessing the possibility

[23] It holds that:

$$D^{max} - [p_0\alpha + (1-p_0)\gamma] = \frac{\lambda^2 p_0(1-p_0)}{(\lambda+\mu+r)(p_0\lambda+\mu+r)}(S-\gamma) > 0 \iff S > \gamma$$

This is intuitive: D^{max} is the value that makes a firm indifferent between disclosing today and postponing disclosure by dt, when the belief is p_0. On the other hand, $(\alpha-\gamma)\,p_0+\gamma$ is the value of D that makes the firm indifferent between disclosing today and never disclosing.

[24] The regulator is always weakly better off under the contract of Proposition 23, and strictly so for most parameter cases.

of concealing evidence regarding the event.

With some abuse of notation, I will be referring to different firm types by their time-zero (or *interim*) beliefs, H, L. In equilibrium, the deterministic law of motion for firm beliefs implies that each type chooses a deterministic stopping time, given the policy in place – where I refer to a firm's *type* as time-zero belief which they have at the beginning of state b.

The principal wants to maximize the expected social welfare $E[W(x, T_{p_0}, p_0)]$, where T_{p_0} is the optimal stopping time chosen by type $p_0 \in \{L, H\}$. Note that x is the single level of ex-ante investment chosen by the firm before the arrival of private information. I similarly define the ex-ante expected continuation value for the firm, $V_b = E[V_b(p_0)]$.

I first describe some properties of $V_b(p_0)$. Then, we will proceed to identify a family of candidate optimal contracts, that minimize V_b for the firm for any given pair of stopping times T_L, T_H that the principal wants to induce. Finally, I will describe the principal's optimal choice amongst the contracts in this optimal family.

4.7.1.1 Single-crossing property

Fix $E(t) = 0$; this is optimal by previous arguments. In line with the previous definition in Equation (4.1), call $U_b(T, D(T), p)$ the expected discounted utility for a firm with current belief p that stops after time T and faces the disclosure payoff $D(T)$. Because $E(t) = 0$, the value U_b only depends on the schedule used through the single value $D(T)$. U_b is always increasing in the disclosure payoff, strictly so if $T < \infty$. Additionally, if $D < \alpha$ (sufficient condition), then the partial derivative with respect to the firm's current belief, $\frac{\partial U_b}{\partial p}$ is positive.[25] By

[25]This is seen by the same argument that leads to Lemma 28. Intuitively, the firm benefits from a private exit if S is sufficiently large relative to D. The condition $D(t) < \alpha$ will always hold for each t in the optimal schedule for reasons previously discussed – even when D is constant, a disclosure payoff of α can induce immediate stopping by even the most optimistic firm. In Proposition 20, $\alpha \geq D^{max}$ and the inequality is strict except if $p_0 = 1$.

revealed preference, this immediately implies that for any contract $\{D(t), E(t)\}$ in place, the optimistic type H will obtain (weakly) greater utility than type L in equilibrium.

It also holds that higher belief p increases the *marginal* benefit from additional experimentation, yielding the following result.

Lemma 34. *: For any delay-dependent schedule $\{D(t), E(t)\}$ that induces stopping times T_H, T_L for the two types and satisfies $D(t) \leq \alpha$ for all t, it must be that $T_H \geq T_L$.*

Thus, it is without loss to restrict attention to contracts that induce stopping times $T_H \geq T_L$ for the two types. At $t = 0$, when privately learning the value $p_0 = p$, the firm that faces schedule $D(t)$ chooses a stopping time T^* that maximizes $V(T, D(T), p)$. Since the path of beliefs conditional on staying is deterministic, as soon as it enters state b, the firm chooses a combination $(T, D(T))$. A point $(t, D(t))$ offers the same utility as point $(t', D(t'))$ if and only if it satisfies:

$$U_b(t, D(t), p_0) = U_b(t', D(t'), p_0) \tag{4.12}$$

The previous Lemma immediately implies a *single-crossing property* for indifference curves in the $T - D$ plane:

Lemma 35. *[Single-Crossing Property] Pick a point $(T, D(T))$, with $D(T) < \alpha$: For any two beliefs $0 < L < H < 1$, it holds that the indifference curve of type L through point $(T, D(T))$ lies **below** that of type H if and only if $t < T$.*

The intuition for the above property is that the optimistic type (H) is less inclined than the pessimistic one to stop early, thus requires *greater* compensation to be indifferent between stopping at T_A and at some $t < T_A$, holding $D(T_A)$ fixed. This happens because the optimistic type believes that receiving the high exit payoff S is more likely. On the other hand, the pessimist needs to be promised a higher

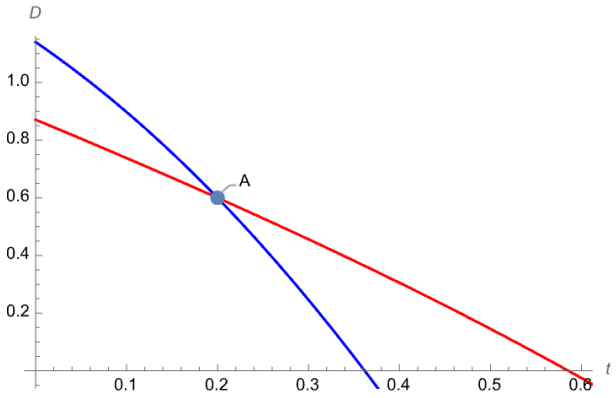


Figure 4.2: Illustration of Lemma 35. The plotted indifference curves of types $L = 0.2$ (red) and $H = 0.7$ (blue) through point $A = (0.2, 0.6)$ satisfy the Single-Crossing Property.

future disclosure payoff than the optimist in order to keep waiting past some time T_A. This is because the pessimist has lower value of hiding than the optimist due to lower perceived chance of dealing with the breach.

4.7.1.2 A family of optimal contracts

Given these preliminary properties of the firm's indifference curves, we can make further progress in narrowing down the set of candidate optimal policies that the principal chooses from. Under Assumptions 1 and 2, the principal's objective W is decreasing in either type's stopping time, and increasing in x. Hence, the principal prefers that V_b is lower, holding the two stopping times fixed. I proceed to identify the family of candidate optimal policies in three steps.

Step 1. I appeal to a version of the Revelation Principle for our setting: For any time-varying contract $C^0 = \{D^0(t), E(t) = 0\}$ that induces stopping times $T_H \geq T_L$ for the two types, the *two-point contract* (2PC), C^1, that uses $D^1(T_L) = D^0(T_L)$, $D^1(T_H) = D^0(T_H)$ and $D^1(t) = 0$ at all other times, induces the same stopping times and yields the same interim utility for each type as the initial contract C^0. It is thus without loss to restrict attention to contracts that offer voluntary disclosure payoffs $D(t) > 0$ at most at two different disclosure delays.[26]

[26] I say "at most" since it might be that $T_H = T_L$ in the initial contract.

This argument allows us to restrict attention to *incentive-compatible*, two-point contracts that induce weakly later stopping times for the H-type. I will only be thinking about contracts with $E(t) = 0$ throughout, hence we can summarize a 2PC by its two points $H = (T_H, D_H)$ and $L = (T_L, D_L)$.

Definition 1. *An incentive-compatible 2PC (IC-2PC) is one that satisfies the two constraints:* $(IC_H) : V_b(H) \geq U_b(T_L, D_L; H)$ *and* $(IC_L) : V_b(L) \geq U_b(T_H, D_H; L)$.

Definition 2. *A 2PC satisfies individual-rationality, if each type weakly prefers his allocation to never disclosing:* $V_b(H) \geq U_b(\infty, H)$ *and* $V_b(L) \geq U_b(\infty, L)$.

The IR curve of type p_0 is always downward sloping and is the locus of points that give this type value equal to their outside option of never disclosing:

$$D(t) = (\alpha - \gamma)\, p(p_0, t) + \gamma \tag{4.13}$$

On the IR curve, and at each time t after the onset of a breach, when the firm has belief $p(p_0, t)$, it is indifferent between disclosing at t or never disclosing.

Step 2. Starting from any incentive-compatible 2PC, we can find a new IC-2PC, with $T_L = 0$, and T_H, D_H same as before, such that type L earns weakly **lower** interim expected utility than under the old contract. As Figure 4.3 graphically shows, this is also an immediate implication of the SCP we have discussed: Since IC_H is above IC_L, we can offer the L type the point at the intersection of the $T = 0$ axis and his indifference curve through point H. So, we can further restrict our attention to contracts that induce *immediate disclosure* by the pessimistic type and that make IC_L bind.

Step 3. Starting from an incentive-compatible 2PC, C^0, with $T_L^0 = 0$ and $\{D_L^0, D_H^0, T_H^0\}$ such that IC_L binds, we can find a new pair of payments $\{D_L^1, D_H^1\}$ such that the two stopping times remain the same, IC_L binds and IR_H binds too. Both type's interim payoff $V_b(p_0)$ is lower as a result.

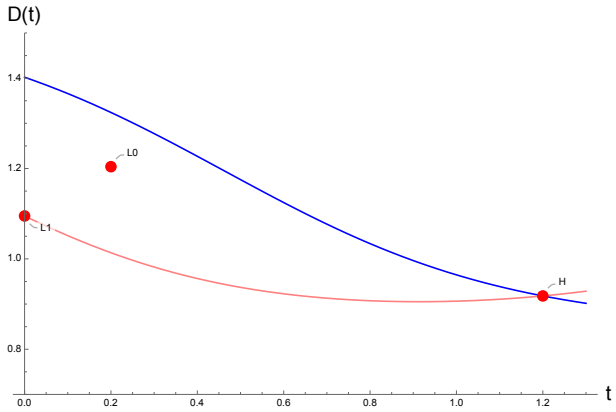


Figure 4.3: Horizontal axis is time since the onset of a breach, t and the vertical is disclosure payoff for stopping at that time, $D(t)$. Starting from 2PC (L_0, H), we can replace L_0 with L_1 and maintain validity of all IR and IC constraints. The two curves are the indifference curves of the two types through point H. The blue curve is the IR_H constraint.

The three steps imply the following Proposition. In the Proposition, I show that we can restrict attention to a family of contracts whose members are indexed by the stopping time they induce for type H, T_H. From now on, I will write $V_b(p_0; T_H)$ as the b-state value of type p_0 given that the contract with index T_H of this family is in place.

Proposition 24. *A regulator whose objective is increasing in investment and decreasing in each type's stopping time will choose a contract from the following family a family of two-point-contracts. Contracts in this family are indexed by the principal's choice of type-H disclosure time, T_H. For every contract in this family:*

- *The pessimistic type discloses immediately, $T_L = 0$.*
- *The optimistic firm earns interim expected utility as if it never discloses, $V_b(H; T_H) = (\alpha - \gamma)H + \gamma$.*
- *Type L earns information rent, $V_b(L; T_H) \geq (\alpha - \gamma)L + \gamma$. The choice of contract T_H determines L type's information rent.*
- *The interim expected discounted value of the L type is higher under contracts with lower T_H.*

The first three points of the Proposition fully characterize the family of contracts from which the regulator will choose. In the third point of Proposition 24, I refer to a type's *information rent* as the difference between $V_b(p_0; T_H)$ of type p_0 and and their *minimum interim* expected utility, which they would receive in the full information case of Proposition 23. We know the latter is given by $(\alpha - \gamma)p_0 + \gamma$. Since for every contract in this family the L type stops immediately and receives $D(0) = U_b(T_H, D_H, L)$, that is exactly that type's interim utility under such a contract. T_H determines the payoff D_H required to have the H type's IR constraint bind. Then, T_H and $D(T_H)$ jointly determine $D(0)$, the payoff that is necessary to have the IC_L constraint bind.

The final point of the above Proposition shows that the principal faces a *trade-off* when choosing a contract from this family. By usual logic, investment incentives depend on the ex-ante continuation payoff V_b. According to the above Proposition, when the principal uses the contract with index T_H, this ex-ante expected utility is given by:

$$\mathbb{E}V_b(T_H) = (1-q)\underbrace{(\beta H + \gamma)}_{=V_b(H;T_H)} + q\, V_b(L; T_H) \tag{4.14}$$

So a trade-off arises for the principal if $V_b(L; T_H) = U_b(T_H, D(T_H), L)$ is *decreasing* in T_H, i.e. if the information rent required to pay type L goes down when we allow H to disclose with greater delay. An increase in T_H has the following two effects on the utility the L type gets from disclosing at $(T_H, D(T_H))$: First, it *postpones* disclosure, holding the disclosure payoff fixed, and second it *decreases* the disclosure payoff required to make IR_H binding, holding disclosure time fixed. The latter effect is always negative and always dominates the first effect, which can sometimes be indeed positive. This is another consequence of single crossing: type H is indifferent and obtains equal $V_b(T_H, H)$ for any schedule T_H in place. When changing to a 2PC with $T'_H > T_H$, the principal *decreases* D to $D(T'_H)$ that keeps H *indifferent* between the old and new schedules. The payment $D(T'_H)$ that

keeps H indifferent between the old and new contracts is not enough to maintain indifference for L, who is more pessimistic and thus *less* inclined that H to extend experimentation up to T'_H. Hence, L becomes worse off at under the contract with higher T_H.

The contract that induces $T_H = 0$ *maximizes* the above ex-ante utility in (4.14), and the contract that induces $T_H = \infty$ minimizes it. The trade off for the regulator is between decreasing the disclosure delay of the high-prior type, versus making it more ex-ante desirable for the firm to avoid a breach.

Relation to canonical screening. Note that although it is the optimistic type that earns higher interim EU from every "allocation" (T, D), it is also that type whose interim IR constraint binds at any optimal contract. In the canonical two-type screening of Mussa and Rosen [1978],[27] it is the high-valuation type who consumes optimally (no distortion at the top) with binding IC and the low type's IR constraint binds, the *opposite* of what Proposition 24 prescribes. This difference can be explained by thinking of the *full-information* case. As we know from the previous section and Proposition 23, under full information each type will disclose immediately and receive $D(0; p_0) = U_b(\infty, p_0) = (\alpha - \gamma)p_0 + \gamma$, which is increasing in p_0. Thus, under full information it is the **pessimistic** type that envies the optimist and is thus that type that needs to earn information rent under private information. In contrast, in the screening model of Mussa and Rosen [1978] it is the high valuation type that envies the low-type's full information allocation.

Implementation. All $D(t)$ schedules used in the family of optimal 2PCs we identified share one unappealing feature: they are **not** (weakly) decreasing in time. This is easy to see, since $D(t) = 0$ for all $t \in (T_L, T_H)$. Perhaps it would be undesirable, or politically infeasible in the case of the principal being a regulator, that the payoff that firms are left with when voluntarily disclosing bad outcomes

[27] Where a seller offers menu of quantity and total payments to buyers with utility $v = \theta q - p$.

is *decreasing* in the delay of disclosing such outcomes.

A question that arises then is whether the equilibrium outcome induced by a given a two-point contract can be replicated using a contract in which $D(t)$ is decreasing in t. Starting from a two-point contract with points $\{(0, D(0)), (T_H, D(T_H))\}$, we want to find a contract that includes both points L, H, and satisfies both IC constraints. IR constraints are automatically satisfied if the initial points are included in the new schedule D'.

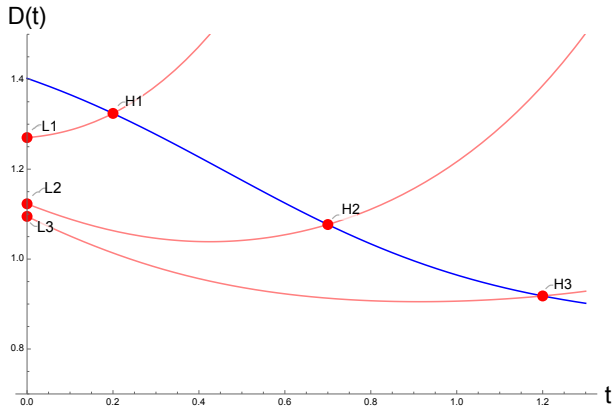


Figure 4.4: The horizontal axis is the time since onset of a breach and the vertical axis is the $D(t)$ schedule. Each pair of points connected by a pink curve (IC_L) is a different IC-2PC from the family of optimal ones. Two observations: (1) Contracts with lower T_H also have higher D_L, confirming the result in Proposition 24, i.e. a trade-off exists. (2) For each contract, the IC_L is *increasing* at the respective T_H.

Unfortunately, this is not always possible. In Figure 4.4, some contracts from the optimal family are presented. For all contracts shown, the indifference curves of the low type going through the high type's allocation are upward sloping at point (T_H, D_H). When this the case, *any* decreasing schedule $D(t)$ (continuous or not) will necessarily violate the IC_L constraint. In Appendix 4.G, we discuss in detail why this increasing segment appears and characterize the region of parameters for which it does. I identify the parameter regions for which IC_L curves are always decreasing, in which case the principal can implement the desired policy by using a decreasing $D(t)$ schedule: the IC_L curve itself.

4.8 Conclusion

I have analysed a model in which a firm invests to avoid data breaches and chooses whether and *when* to disclose data breaches that occur. The regulator, who wishes to incentivize both ex-ante investment and ex-post prompt disclosure, faces a trade-off when choosing the penalty (or level of support) for firms that voluntarily disclose breaches. I formally analyzed the firm's disclosure decision as an optimal stopping problem, and showed how the regulator's optimal policy is shaped by payoff parameters, the way in which the firm learns about the data breach while hiding, and, importantly, on whether the regulator can ex-post verify the delay with which a firm voluntarily discloses a breach.

There are further questions to explore, especially building on the extension of Section 4.6.1, which relaxed the assumption that the firm learns about a data breach immediately. I studied the impact of such a change when the regulator cannot write contracts on the delay of disclosure. If the regulator cannot verify when the firm got knowledge of the breach but *can* verify the time the breach occurred (even if the firm did not find out at the time), the simple deadline contracts presented in Section 4.7 do not work and neither do the optimal two-point contracts identified later: a firm that realizes the breach even dt after it happened, might optimally never disclose under these contracts. Exploring the regulator's choices in that setting is a particularly intriguing direction for future research.

Relatedly, in Section 4.6.1, the firm (or the regulator) cannot influence the delay with which it becomes aware of a breach. Extending the model to account for endogenous *monitoring* of data breaches will further enrich both the predictions that the model produces, but also introduce an additional important dimension to the regulator's problem of incentive provision.

Appendix 4.A Analysis

Proof of Proposition 19

As already argued in the main text, the deterministic path of beliefs about θ implies that the firm can be thought of as choosing a *deterministic* stopping time $T \in [0, \infty)$, which is the voluntary disclosure time, conditional on no exit up to that time. For a proof deriving the value function via the optimality principle and solving the HJB equation, see Appendix 4.F. I characterize the shape of U_b, defined in equation (4.1), and the optimal disclosure time as a function of the model's parameters. Computing the integral yields:

$$\begin{aligned} U_b = p_0 \overbrace{\frac{\pi_b + \lambda S + \mu E}{\lambda + \mu + r}}^{=a(E)} \left(1 - e^{-(\lambda+\mu+r)T}\right) + p_0 e^{-(\lambda+\mu+r)T} D \\ + (1 - p_0) \underbrace{\frac{\pi_b + \mu E}{\mu + r}}_{=\gamma(E)} \left(1 - e^{-(\mu+r)T}\right) + (1 - p_0) e^{-(\mu+r)T} D \end{aligned} \tag{4.15}$$

and using the definitions of $\alpha(E)$ and $\gamma(E)$ from the main text this is more concisely:

$$U_b = p_0\, \alpha(E) + p_0 e^{-(\lambda+\mu+r)T}(D - \alpha) + (1 - p_0)\gamma + (1 - p_0) e^{-(\mu+r)T}(D - \gamma(E))$$

and the derivative with respect to the stopping time T is:

$$\frac{\partial U_b}{\partial T} = -p_0\, e^{-(\lambda+\mu+r)T^*}(\lambda + \mu + r)(D - \alpha(E)) - (1 - p_0)(\mu + r) e^{-(\mu+r)T^*}(D - \gamma(E))$$

From the above, we already conclude that: (1) if $D < \min\{S, \gamma(E)\}$, then $U_b'(T) > 0$ for all T, so $\hat{T} = \infty$, i.e. voluntary disclosure is never optimal and (2) if $D > \max\{S, \gamma(E)\}$ then $U_b'(T) < 0$ for all T and $\hat{T} = 0$. Continuing is never optimal, since disclosure is the optimal mode of exit. There are two remaining

cases to consider: $S > D > \gamma(E)$ and $\gamma(E) > D > S$. The derivative is positive if and only if:

$$\big[(\pi_b + \lambda S) - (\lambda + \mu + r)D\big] + \omega(p_0)e^{\lambda T}\big(\pi_b - (\mu + r)D\big) > 0$$

where I defined $\omega(p) := (1-p)/p$.

Case 1. $S > D > \gamma(E)$

For the case of $D > \gamma(E)$ (regardless of S):

$$\begin{aligned}
&\frac{\partial U_b}{\partial T} > 0 \iff \\
&e^{\lambda T} < \frac{1}{\omega(p_0)} \frac{\big[(\pi_b + \lambda S) - (\lambda + \mu + r)D\big]}{-\big(\pi_b - (\mu + r)D\big)} \\
&\quad = \frac{1}{\omega(p_0)} \frac{\lambda(S - D) - (\mu + r)(D - \gamma(E))}{(\mu + r)\big(D - \gamma(E)\big)} \\
&\quad = \frac{1}{\omega(p_0)} \left(\frac{1}{p^*} - 1\right) = \frac{\omega(p^*)}{\omega(p_0)}
\end{aligned}$$

where I have defined $p^* := (\mu + r)(D - \gamma(E))/[\lambda(S - D)] = \kappa(D - \gamma(E))/(S - D)$, for $\kappa := (\mu + r)/\lambda$. This shows that if $D > \gamma(E)$, then $U_b(T)$ is quasi-concave in T. The maximizer is $\hat{T} = (1/\lambda) log(\omega(p^*)/\omega(p_0))$, positive if and only if $p^* < p_0$. If $p^* < p_0$, U_b is maximized at $\hat{T} = 0$. It follows that if $D > \gamma(E)$, then the firm discloses in finite time, regardless of the value of S. This is intuitive since for sufficiently long exploration without a breakthrough, $p_0 \to 0$, so S does not matter.

Case 2. $\gamma(E) > D > S$.

Identical algebra to that above reveals that U_b is quasi-convex in T, hence the solution to the firm's first-order condition yields the *minimizer* of U_b. The firm chooses between the two extremes $\hat{T} \in \{0, \infty\}$, i.e., compares $U_b(0; D, E) = D$ with $U_b(\infty; D, E) = p_0\,\alpha + (1 - p_0)\gamma(E)$. It is optimal to never disclose if:

$$[\omega(p_0)(k + 1) + k](\gamma(E) - D) - (D - S) > 0$$

If λ is large and $D >> S$, it may be optimal to disclose for large values of p to **avoid** an exit to S. But for $p \to 0$, it must be that continuing forever without disclosure is optimal. The value of the above difference *decreases* in p_0. As $p_0 \to 0$, $\omega(p_0) \to \infty$ and the inequality is true, confirming this intuition.

4.A.1 Proof of Lemma 22

The threshold in terms of the posterior belief is computed in the proof of Proposition 19.

4.A.2 Proof of Lemma 24

At the beginning of the game, the firm chooses $x = 1$ if the expected discounted present value of doing so is greater than under $x = 0$:

$$-C + \int_0^\infty e^{-rt} e^{-h_1 t}(\pi_s + h_1 V_b(D,E))dt > \int_0^\infty e^{-rt} e^{-h_0 t}(\pi_s + h_0 V_b(D,E))dt \iff$$

$$-C + \frac{\pi_s + h_1 V_b(D,E)}{r + h_1} > \frac{\pi_s + h_0 V_b(D,E)}{r + h_0} \iff \tag{4.16}$$

$$V_b(D,E) < \frac{1}{r}\left[\pi_s - \frac{C}{h_0 - h_1}(r + h_1)(r + h_0)\right] := V^{max} \tag{4.17}$$

4.A.3 Proof of Lemma 25

Given $D^w > \max\{S^w, \gamma^w\}$, reasoning identical to that of the proof of Proposition 19 reveals that the regulator-optimal disclosure time is $T^{FB} = 0$ and that W_b is decreasing in T. Note that by the logic of the same proof, this is a sufficient and not necessary condition for $T^{FB} = 0$. Next, I ask, holding T fixed, when the regulator prefers $x = 1$ to be played. Under $D^w > \max\{S^w, \gamma^w\}$, the condition is *strictest* for $T = 0$, when W_b takes its largest value, $W_b(0) = -L_0 + D^w$, where L_0 is the instantaneous loss to the regulator at the onset of a breach, and D^w is the

termination payoff of the regulator following disclosure of a breach. The regulator prefers $x = 1$ also if the cost of security is low relative to the loss of entering state b. Following the reasoning of Lemma 24, there is a value $W^{max}(w_s, h_0, h_1, r, C)$ such that when the firm uses $T = 0$, the regulator prefers $x = 1$ if and only if $W_b(0) \leq W^{max}(w_s, h_0, h_1, r, C)$, i.e. if and only if:

$$-L_0 + D^w \leq \frac{1}{r}\left[w_s - \frac{C}{h_0 - h_1}(r + h_1)(r + h_0)\right] \iff$$

$$\frac{C}{h_0 - h_1}(r + h_1)(r + h_0) \leq r\left(L_0 - D^w\right) + w_s \iff \tag{4.18}$$

$$r\left(D^w - L_0\right) \leq w_s - \frac{(r + h_1)(r + h_0)}{h_0 - h_1}C \tag{4.19}$$

Observe that this holds for sufficiently large L_0. An arbitrarily large value of D^w relative to the instantaneous cost of a breach means that the regulator has *weaker* investment incentive than the firm when $T = 0$, which could be micro-founded through a motive for **experimentation**; let a firm be breached and disclose the risk for society to benefit at large. I abstract from this story in order to focus on the case where a trade-off exists.

4.A.4 Proof of Proposition 20

The arguments stated in previous lemmas make it clear that $E^* = 0$ under Assumptions 1 and 2. A higher value of E would mean that (1) $p(D, E)$ decreases, i.e. disclosure is delayed, and also that $V_b(D, E)$ increases, i.e. that IC_x becomes harder to satisfy. So $E^* = 0$.

Regarding D, using the definition of $p(D, E)$ in (4.5), I define as D^{min} the largest value of D such that $T = \infty$ or equivalently, $p^* \leq 0$. Given $S > \gamma$, that value is $D^{min} = \gamma$. When $T = \infty$ is played, the firm receives expected value in state b of $U_b(\infty, D) = \alpha p_0 + \gamma(1 - p_0)$, which is independent of D since disclosure never occurs, hence $V_b(D^{min}) = \alpha p_0 + \gamma(1 - p_0)$. That is the *lowest* that the firm's

state-b continuation value can be in equilibrium. If $V_b(D^{min}) > V^{max}$, the threat of a breach is never enough to incentivize the firm to invest in security, hence the regulator can never achieve $x^* = 1$. The optimal policy becomes to set D at any value large enough to induce $T = 0$.

To induce an earlier stopping time, the regulator must increase D beyond the value of $D^{min} = \gamma$. Looking at the expression for $p(D, E)$, the *lowest* value of D that induces immediate disclosure, i.e. the one that satisfies $p^*(D, 0) = p_0$:

$$\begin{aligned} p^*(D,0) = p_0 &\iff \\ \frac{(\mu+r)(D-\gamma)}{\lambda(S-D)} = p_0 &\iff \\ D = \frac{(\mu+r)\gamma + p_0\lambda S}{\mu + r + \lambda p_0} &:= D^{max} \end{aligned}$$

Observe that $D^{max} \to \alpha$ as $p_0 \to 1$: when the firm is certain that $\theta = 1$, it either discloses immediately or never and that depends on whether $D > \alpha$. For $p_0 < 1$, if D^{max} is used, then the regulator induces immediate disclosure, and does so by inducing the *lowest* value of V_b consistent with $T(D) = 0$. It is clear that if that value is lower than V^{max}, then IC_x is satisfied and the first-best achieved.

In the case of $D^{min} < V^{max} < D^{max}$, and by Lemma 23, the monotonicity of $V_b(D)$ implies that there is a unique value of $D \in (D^{min}, D^{max})$ that yields $V_b(D^*) = V^{max}$, and the associated disclosure time is positive $T(D^*) = T^*$.

Appendix 4.B Comparative Statics

4.B.1 Direct effects on W_b

For all subsequent calculations, define the following quantities, directly analogous to those of the firm's problem.

$$\gamma^w := \frac{w_b + \mu E^w}{\mu + r}$$

and

$$\alpha^w := \frac{\lambda S^w + w_b + \mu E^w}{\mu + r + \lambda} = \frac{\lambda S^w + (\mu + r)\gamma^w}{\lambda + \mu + r}$$

so that in direct analogy to U_b:

$$W_b = p_0\Big(\alpha^w + e^{-(\lambda+\mu+r)T}(D^w - \alpha^w)\Big) + (1-p_0)\Big(\gamma^w + e^{-(\mu+r)T}(D^w - \gamma^w)\Big) \quad (4.20)$$

Lemma 27: Effect of μ on W_b

For this calculation, I am explicit about the dependence of α^w and γ^w on the parameter μ.

$$\frac{\partial W_b}{\partial \mu} = p_0\Big(\alpha'^w(\mu)(1 - e^{-(\lambda+\mu+r)T}) - Te^{-(\lambda+\mu+r)T^*}(D^w - \alpha^w(\mu))\Big) + (1-p_0)\Big(\gamma'^w(\mu)(1 - e^{-(\mu+r)T}) - Te^{-(\mu+r)T}(D^w - \gamma^w(\mu))\Big)$$

This is zero at $T = 0$. Differentiating again with respect to the firm's disclosure time, T yields:

$$\frac{\partial^2 W_b}{\partial T \partial \mu} > 0 \iff (E^w - D^w) + T\Big[\lambda(D^w - S^w) + (\mu + r)(D^w - \gamma^w)\Big] > 0$$

which is true under the assumption $D^w > \max\{S^w, \gamma^w\}$ and for E^w sufficiently close to D^w (either larger or smaller.

Lemma 28: Effect of p_0 on W_b

Partially differentiate (4.20) with respect to p_0 to get:

$$\frac{\partial W_b}{\partial p_0} = (a^w - \gamma^w) + e^{-(\lambda+\mu+r)T}(D^w - \alpha^w) - e^{-(\mu+r)T}(D^w - \gamma^w)$$

which is zero at $T = 0$. The mixed partial derivative is given by:

$$\frac{\partial^2 W_b}{\partial p_0 \partial T} = -(\mu + r + \lambda)e^{-(\lambda+\mu+r)T}(D^w - \alpha^w) + (\mu + r)(D^w - \gamma^w)e^{-(\mu+r)T} > 0 \iff$$

$$\text{using } \kappa := (\mu + r)/\lambda, \quad -(\frac{1}{\kappa} + 1)(D^w - \alpha^w) + e^{\lambda T}(D^w - \gamma^w) > 0 \iff$$

$$(\frac{1}{\kappa} + 1)(D^w - \alpha^w) < e^{\lambda T}(D^w - \gamma^w) \iff$$

$$e^{\lambda T} > (\frac{1}{\kappa} + 1)\frac{(D^w - \alpha^w)}{(D^w - \gamma^w)}$$

If $\alpha^w > D^w$, this is satisfied for any $T \geq 0$. Otherwise, this is positive iff $T > \frac{1}{\lambda} log\left[(\frac{1}{\kappa} + 1)\frac{(D^w - \alpha^w)}{(D^w - \gamma^w)}\right]$ and $\partial W_b/\partial p_0$ is a quasi-convex function of p_0. The limit as $T \to \infty$ is:

$$\lim_{T\to\infty} \frac{\partial W_b}{\partial p_0} = a^w - \gamma^w > 0 \iff S^w > \gamma^w$$

Lemma 29: Effect of λ on W_b

Partially differentiate (4.20) to get:

$$\frac{\partial W_b}{\partial \lambda} = p_0\Big(\alpha'^w(\lambda)(1 - e^{-(\lambda+\mu+r)T}) - Te^{-(\lambda+\mu+r)T}(D^w - \alpha^w(\lambda))\Big)$$

This is zero at $T = 0$ and:

$$\lim_{T\to\infty} \frac{\partial W_b}{\partial \lambda} = \frac{\partial a^w}{\partial \lambda} = \frac{(\mu + r)(S^w - \gamma^w)}{(\mu + r + \lambda)^2} > 0 \iff S^w > \gamma^w$$

Partially differentiating again with respect to T yields:

$$\frac{\partial^2 W_b}{\partial T \partial \lambda} = e^{-(\lambda+\mu+r)T^*}\Big[(\lambda + \mu + r)\,\alpha'^w(\lambda) + T(\lambda + \mu + r)(D - \alpha^w(\lambda)) - (D - \alpha^w(\lambda))\Big]$$

which is positive iff:

$$(\lambda + \mu + r)\,\frac{(\mu + r)(S^w - \gamma^w)}{(\mu + r + \lambda)^2} + T(\lambda + \mu + r)(D^w - \alpha^w(\lambda)) - (D^w - \alpha^w(\lambda)) > 0 \iff$$

$$(\mu + r)(S^w - \gamma^w) + T(\lambda + \mu + r)^2(D^w - \alpha^w(\lambda)) - (\lambda + \mu + r)(D^w - \alpha^w(\lambda)) > 0 \iff$$

$$T(\lambda + \mu + r)^2(D^w - \alpha^w(\lambda)) > (\lambda + \mu + r)(D^w - \alpha^w(\lambda)) - (\mu + r)(S^w - \gamma^w) > 0 \iff$$

$$T(\lambda + \mu + r)^2(D^w - \alpha^w(\lambda)) > (\lambda + \mu + r)(D^w - S^w) \iff$$

$$T\big[\lambda(D^w - S^w) + (\mu + r)(D^w - \gamma^w)\big] > (D^w - S^w)$$

Under Assumption 1, both sides are positive, hence there is a unique T' such that $\partial W_b/\partial\lambda > 0$ iff $T > T'$, in other words, $\partial W_b/\partial\lambda > 0$ is a quasi-convex function of T. Since it is zero at $T = 0$, it becomes positive only if $\lim_{T\to\infty} \frac{\partial W_b}{\partial \lambda} > 0 \iff S^w > \gamma^w$.

4.B.2 Proof of Lemma 30

The expression for the stopping time as a function of λ is:

$$T(\lambda) = \frac{1}{\lambda} log\left(\frac{\omega(p(\lambda))}{\omega(p_0)}\right) = \frac{1}{\lambda}\Big[log(\omega(p(\lambda)) - log(\omega(p_0))\Big] \tag{4.21}$$

where I have written the posterior stopping threshold as a function of λ. We can write $p(\lambda) = A/\lambda$, with $A > 0$, and $\omega(p) = (1 - p)/p$, we can write $\omega(p(\lambda)) =$

$(\lambda/A - 1)$. Differentiating with respect to λ yields:

$$\begin{aligned} T'(\lambda) &= \frac{(-1)}{\lambda^2} log\left(\frac{\omega(p(\lambda))}{\omega(p_0)}\right) + \frac{1}{\lambda}\frac{1}{\omega(p(\lambda)}\frac{1}{A} \\ &= \frac{(-1)}{\lambda} T(\lambda) + \frac{1}{\lambda}\frac{1}{(\lambda - A)} \\ &= \frac{1}{\lambda}\Big[T(\lambda) - \frac{1}{\lambda}\frac{1}{(\lambda - A)}\Big] \end{aligned} \tag{4.22}$$

In the second step, I used the definition of $T(\lambda)$. To verify quasi-concavity, note that the second derivative is always negative at points that satisfy the first-order condition, $T'(\lambda) = 0$.

$$\begin{aligned} T''(\lambda) &= \frac{-1}{\lambda} T'(\lambda) + \frac{1}{\lambda^2}\Big[T(\lambda) - \frac{1}{\lambda}\frac{1}{(\lambda - A)}\Big] + \frac{1}{\lambda}\frac{(-1)}{(\lambda - A)^2} \\ &= \frac{1}{\lambda}\frac{(-1)}{(\lambda - A)^2} < 0 \end{aligned}$$

where for the last equality I have used the first-order condition to drop both the first and second terms of the sum. To prove the second part of the result, I take the limit of $T'(\lambda)$ as λ goes to ∞, keeping in mind that $\lim_{\lambda\to\infty} p(\lambda) = 0$, and the limit as λ approaches the value that solves $p(\lambda) = p_0$ and thus $T(\lambda) = 0$; call that value λ_0. Taking the limit as $\lambda \to \lambda_0$ using the expression in (4.22) is straightforward and yields $\lim_{\lambda\to\lambda_0} T'(\lambda) = K_0 > 0$. The limit as λ approaches ∞ is:

$$\begin{aligned} \lim_{\lambda\to\infty} T'(\lambda) &= \lim_{\lambda\to\infty}\left[\frac{(-1)}{\lambda^2}[log(\omega(p(\lambda)) - log(\omega(p_0))] + \frac{1}{\lambda}\frac{1}{(\lambda - A)}\right] \\ &= -\lim_{\lambda\to\infty}\frac{log(\omega(p(\lambda))}{\lambda^2} \\ &= -\lim_{\lambda\to\infty}\frac{log(\lambda - A) - log(A)}{\lambda^2} \\ &= -\lim_{\lambda\to\infty}\frac{log(\lambda - A)}{\lambda^2} \\ &= 0 \end{aligned}$$

To obtain the final limit, we applied L'Hôpital's rule.

4.B.3 Proof of Lemma 31

Part (1)

I study the total effect of changes in parameters in T^{min}, the earliest stopping time that the regulator can implement, subject to IC_x holding, and I focus on cases in which $T^{min} > 0$. Taking π_b as an example, total differentiation shows:

$$\frac{dT^{min}}{d\pi_b} = \underbrace{\frac{\partial T}{\partial D}}_{(-)} \frac{\partial D^*}{\partial \pi_b} + \underbrace{\frac{\partial T}{\partial \pi_b}}_{(+)} \tag{4.23}$$

There is a *direct* effect on the firm's preferred stopping time, and an *indirect* one via the regulator's adjustment of D^*. In this case, the direct one is positive, since π_b induces further experimentation, $\partial p(D)/\partial \pi_b < 0$. For the indirect effect, I remind the reader that D^* is the unique value of the disclosure payoff such that $T(D^*) = T^*$, and $V_b(D^*) = V^{max}$. Thus, to sign the total derivative in (4.23), we need to understand how the regulator changes D^* in response to a change in the relevant parameter. I use the definition of D^* as the unique value of D that makes the ex-ante incentive compatibility constraint, IC_x, *bind*.

$$V^{max} = V_b(\pi_b, D^*(\pi_b))$$

where I am being explicit about the dependence of the regulator's policy D^* and of the continuation value on π_b. Note that V^{max} does not depend on π_b, so this analysis will apply to all parameters that do not affect V^{max}. Following a change in π_b, the principal will optimally change D^* to maintain IC_x validity:

$$\frac{dV_b}{d\pi_b} = 0 = \frac{\partial V_b}{\partial \pi_b} + \frac{\partial V_b}{\partial D}\frac{\partial D^*}{\partial \pi_b}$$

There is no first-order effect on V_b via the choice of T, hence:

$$\frac{dD^*}{d\pi_b} = -\frac{\partial V_b/\partial \pi_b}{\partial V_b/\partial D^*} < 0 \tag{4.24}$$

The intuition is that since an increase in π_b increases the continuation value V_b, the regulator must respond by *decreasing* D^* in order to maintain validity of IC_x. Thus, T^* will increase through two channels. The direct effect of π_b, holding the policy fixed is to delay voluntary disclosure, and the indirect effect of forcing the regulator to reduce the disclosure payoff further increases the delay with which the breach is voluntarily disclosed. The results for p_0 and μ are obtained through the same steps.

Part (2)

For the parameters that affect V^{max} *without* affecting $V_b(D)$, for example π_s (but also h_0, h_1, and C):
I remind the reader of the V^{max} definition, which is the maximum level of V_b such that the firm invests in security.

$$V^{max} = \frac{1}{r}\left[\pi_s - \frac{(r+h_0)(r+h_1)}{h_0 - h_1}C\right] \iff$$
$$(h_0 - h_1)(\pi_s - rV^{max}) = (r+h_0)(r+h_1)C$$

Parameters like π_s that do not affect state-b payoffs of the firm have no direct effect on the optimal stopping time. But they do affect the level of V^{max} that the regulator can provide to the firm without jeopardizing IC_x.

$$\frac{dT}{d\pi_s} = \underbrace{\frac{\partial T}{\partial D}}_{(-)}\frac{\partial D^*}{\partial \pi_s}$$

where D^* is defined by $V_b(D^*) = V^{max}(\pi_s)$, so the above sign depends on:

$$\frac{\partial D^*}{\partial \pi_s} = \frac{\partial V^{max}/\partial \pi_s}{\partial V_b/\partial D} > 0$$

Earlier disclosure will be achievable as a result of an increase in a parameter that increases the value of V^{max}. Increasing a parameter that increases V^{max}, means that the firm is more inclined to invest in security, hence **earlier** disclosure is achievable by the D^* policy.

Part (3)

The impact of λ on T^{min} is through a positive indirect effect via D^* and a direct effect, holding the regulator's policy fixed. Since λ affects V_b but not V^{max}, I follow similar reasoning to that of part (1).

$$\frac{\partial D^*}{\partial \lambda} = -\frac{\partial V_b/\partial \lambda}{\partial V_b/\partial D^*} < 0$$

because both numerator and denominator are positive. To sign the direct effect of the breakthrough rate on T^{min}, I appeal to the result of Lemma 30. For any given value of λ, the regulator responds with $D^*(\lambda)$. Given this value of D^*, there exists a largest value of λ, λ_0, such that $T(D^*) = 0$. This is the value that makes the posterior stopping threshold exactly equal to the prior, i.e.:

$$\lambda_0 = \frac{(\mu + r)\, D^*(\lambda) - \pi_b}{(S - D^*(\lambda))\, p_0} \tag{4.25}$$

I focus on regions where $\lambda > \lambda_0$ so that T^{min} is positive. According to Lemma 30, if $\lambda - \lambda_0$ is sufficiently small, an increase in λ will cause an increase in the agent's optimal stopping time, holding D fixed at the initial $D^*(\lambda)$. This signs the direct

effect.

$$\frac{dT^{min}}{d\lambda} = \underbrace{\frac{\partial T}{\partial D}\frac{\partial D^*}{\partial \lambda}}_{(+)} + \underbrace{\frac{\partial T}{\partial \lambda}}_{(+)} > 0 \tag{4.26}$$

4.B.4 Proof of Lemma 32

I remind the reader of the definition:

$$\Delta W(T^{min}) := -C + \frac{1}{(h_0+r)(r+h_1)}\Big[w_s(h_0-h_1) + h_1(r+h_0)W_b(T^{min}) - h_0(r+h_1)W_b(0)\Big] \tag{4.27}$$

Partial differentiation yields:

$$\frac{\partial \Delta W(T^{min})}{\partial h_1} = \frac{-w_s}{(r+h_1)^2} + \frac{rW_b(T^{min})}{(r+h_1)^2} = \frac{-(w_s - rW_b(T^{min}))}{(r+h_1)^2} < 0$$

and

$$\frac{\partial \Delta W(T^{min})}{\partial h_0} = \frac{w_s}{(h_0+r)^2} - \frac{rW_b(0)}{(h_0+r)^2} = \frac{w_s - rW_b(0)}{(h_0+r)^2} > 0$$

where the signs are implied by Lemma 25 and Assumption 2. Given $W_b(T)$ is decreasing in the firm's disclosure time, combining the above direct effects with the results in Lemma 31, we obtain the results.

4.B.5 Proof of Lemma 33

$$\Delta W(T^{min}) = W(T^{min}, 1) - W(0, 0)$$

The results follow immediately from combining the results on the direct effects and the effects via T^{min}. Looking at equation (4.27), both parameters only affect the difference via their impact on $W(T^{min}, 1)$, which is both direct and also indirect via T^{min}.

First, we know that increases in the prior probability p_0 always increase T^{min} from

Lemma 31. The direct effect on welfare $\Delta W(T^{min})$ is directly signed by Lemma 28, since p_0 does not affect the regulator's welfare when $T^* = 0$. By that Lemma, the direct effect can only be positive if T^{min} is large **and** $\alpha^w > \gamma^w$. For low initial value of T^{min}, both the direct and indirect effects of p_0 are negative regardless of $\text{sgn}(\alpha^w - \gamma^w)$.

Second, for the breakthrough rate λ, we know from the proof of Lemma 31 that if λ is sufficiently small, the effect on T^{min} is positive. Additionally, the value of λ has no direct impact on welfare when disclosure happens very soon. Combining these statements, when λ is sufficiently small (i.e. $\lambda \to \lambda_0$, see proof of Lemma 30, part 3), $T^{min} \to 0$ and the dominant effect on welfare is the negative impact of an increase in the stopping time T^{min}.

Appendix 4.C Extensions

4.C.1 Proof of Proposition 21

To see how the relative benefit of inducing investment depends on d, first differentiate ΔW, the difference in the regulator's welfare between the two candidate optimal contracts, defined in equation (4.8), holding T^{min} fixed. Remember that:

$$\begin{aligned} W_b(T,d) =& (1 - e^{-(\mu+r)d})\gamma^w + e^{-(\mu+r)d} W(T,0) \\ =& \gamma^w + e^{-(\mu+r)d}(W(T,0) - \gamma^w) \end{aligned}$$

where by $W_b(T,d)$, I denote the regulator's continuation value of state b when the firm becomes aware with delay d and stops time T *after* realization. When W_b only has one argument, it is implied that $d = 0$, as in the baseline. Hence, the

partial derivative in question is:

$$\begin{aligned}
&\frac{\partial(W(1,T^{min})-W(0,0))}{\partial d}\\
&=\nu\frac{h_1}{r+h_1}(\mu+r)e^{-(\mu+r)d}\big(\gamma^w-W_b(T^{min})\big)-\nu\frac{h_0}{r+h_0}(\mu+r)e^{-(\mu+r)d}\big(\gamma^w-W_b(0)\big)\\
&=\nu\,e^{-(\mu+r)d}(\mu+r)\left[\frac{h_1}{r+h_1}\big(\gamma^w-W_b(T^{min})\big)-\frac{h_0}{r+h_0}\big(\gamma^w-W_b(0)\big)\right]\\
&=\nu\,e^{-(\mu+r)d}(\mu+r)\left[\frac{h_0}{r+h_0}\big(W_b(0)-\gamma^w\big)-\frac{h_1}{r+h_1}\big(W_b(T^{min})-\gamma^w\big)\right]>0\iff\\
&\frac{\frac{h_0}{r+h_0}}{\frac{h_1}{r+h_1}}\big(W_b(0)-\gamma^w\big)-\big(W_b(T^{min})-\gamma^w\big)>0
\end{aligned}$$

and we obtain the positive sign because (1) $h_0/(r+h_0)>h_1/(r+h_1)\iff h_0>h_1$, which is the assumption, (2) $W_b(0)=D^w>\gamma^w$ under Assumption 1, and (3) $W_b(T)\leq W_b(0)$, for all T, under Assumption 1.Hence the direct effect on ΔW is positive. To obtain the direct effect of ν on ΔW, observe that:

$$\begin{aligned}
&\frac{\partial(W(1,T^{min})-W(0,0))}{\partial\nu}\\
&=\frac{h_1}{r+h_1}\big(W_b(T^{min},d)-W_b(T^{min},0)\big)-\frac{h_0}{r+h_0}\big(W_b(0,d)-W_b(0,0)\big)
\end{aligned}$$

Hence, the above derivative is equal to:

$$\begin{aligned}
&\frac{h_1}{r+h_1}(1-e^{-(\mu+r)d})\big(\gamma^w-W_b(T^{min})\big)-\frac{h_0}{r+h_0}(1-e^{-(\mu+r)d})\big(\gamma^w-W_b(0)\big)\\
=&(1-e^{-(\mu+r)d})\left[\frac{h_0}{r+h_0}\big(W_b(0)-\gamma^w\big)-\frac{h_1}{r+h_1}\big(W_b(T^{min})-\gamma^w\big)\right]>0
\end{aligned}$$

and we obtain the positive sign by the argument used to sign the previous derivative.

Next, I show that T^{min} is decreasing in either ν or d. Note that below, I will denote by $V_b(d)$ the continuation value of entering state b for the firm, given that it will become aware of the breach with delay d. Similar to above for the regulator, I will

denote with $\mathbb{E}V_b$ the expectation of V_b with respect to the awareness delay. Similar to how we proceeded in the proofs in the Comparative Statics section, notice that (1) neither ν or d have a direct effect on T^{min}: by assumption, once the firm realizes that a breach has occurred, it faces the same stopping problem regardless of when the breach occurred, (2) both induce a reduction in V_b, and hence (3) an increase in either allows the principal to increase D^*, while maintaining IC_x. Similar to equation (4.24), we obtain:

$$\frac{dD^*}{d(d)} = -\underbrace{\frac{\partial \mathbb{E}V_b/\partial d}{\partial \mathbb{E}V_b/\partial D^*}}_{(+)} > 0 \tag{4.28}$$

and we obtain the first sign via:

$$\begin{aligned}
\frac{\partial \mathbb{E}V_b}{\partial d} &= \frac{\partial}{\partial d}\Big[(1-\nu)\,V_b(0) + \nu\,V_b(d)\Big] \\
&= \frac{\partial}{\partial d}\Big[(1-\nu)V_b(0) + \nu\big[(1-e^{-(\mu+r)d})\gamma + e^{-(\mu+r)d}V_b(0)\big]\Big] \\
&= \nu\,e^{-(\mu+r)d}\big(V_b(0)-\gamma\big) < 0
\end{aligned}$$

since by usual arguments, $\gamma := \pi_b/(\mu+r)$ is the lowest value V_b could ever take.

4.C.2 Proof of Proposition 22

I begin this section by proving that the agent's optimal policy for the case of $\mu_1 = 0$ and $0 \leq \lambda < \mu_0$ is the one I describe in the main text. First, I define the function:

$$J(p) = p\,\hat{\alpha} + (1-p)\,\hat{\gamma} \tag{4.29}$$

where under $E = 0$, the constants are defined as:

$$\hat{\alpha} := \frac{(\pi_b + \lambda S)}{(r + \lambda)} \tag{4.30}$$

$$\hat{\gamma} := \frac{\pi_b}{(\mu + r_0)} \tag{4.31}$$

Under $\mu_0 > \lambda$, it holds that $\hat{\alpha} > \hat{\gamma}$, hence $J'(p) > 0$. The agent's value from never disclosing is increasing in the probability that $\theta = 1$. Next, for given D I define $\hat{p}(D)$ as the unique belief that satisfies $J(p) = D$.

$$J(p) = \hat{D} \iff \hat{p}(D) = \frac{D - \hat{\gamma}}{\hat{\alpha} - \hat{\gamma}} \tag{4.32}$$

Lemma 36. *Given a value of $D \in (\hat{\alpha}, \hat{\gamma})$ and $E = 0$, the agent's optimal policy is to disclose immediately if $p_0 < \hat{p}(D)$; otherwise the agent never discloses.*

Proof. It is simple to see that under $\hat{\alpha} > D > \hat{\gamma}$, the agent stops for beliefs close to zero and continues for beliefs close to $p = 1$. To verify that the candidate policy is indeed the optimal, I will show that the induced value function satisfies the Hamilton-Jacobi-Bellman equation. In particular, I will show that:

1. Continuing is optimal at the region $p \geq \hat{p}$.
2. Stopping is optimal at the region $p \leq \hat{p}$.

Continuing is optimal at the region $p \geq \hat{p}$.

This is straightforward to verify. Under the suggested policy, the value function at any $p > \hat{p}$ is equal to $J(p)$ and by the monotonicity of J, we obtain $J(p) > J(\hat{p}) = D$, thus no deviation is profitable at any belief in the non-disclosure region.

Stopping is optimal at the region $p \leq \hat{p}$.

The induced value function is not differentiable at $\hat{p}$ (it is continuous though). Nevertheless, we can apply standard verification arguments, by taking cases:

Case 1: $p + dp < \hat{p}$.

In this case, if the firm deviates to continuing for dt, it stops in the absence of breakthrough or exposure. The deviation yields payoff:

$$\begin{aligned}
&\pi_b\,dt + e^{-r\,dt}\mathbb{E}[V(p+dp)|p] \\
\simeq &\pi_b\,dt + (1 - r\,dt)\Big[p\,dt\,\lambda\,S + [1 - p\,\lambda\,dt - (1-p)\,\mu\,dt)]D\Big] \\
\simeq &\pi_b\,dt + p\,dt\,\lambda\,S + [1 - p\,\lambda\,dt - (1-p)\,\mu\,dt)]D - r\,dt\,D
\end{aligned}$$

I have used the first-order approximation of $e^{r\,dt}$ and ignored second-order dt terms. This deviation payoff is *smaller* than the on-path payoff of D iff:

$$p < \frac{(\mu + r)D - \pi_b}{\lambda(S - D) + \mu D}$$

and it is simple to verify that the right-hand side is *larger* than $\hat{p}$ under our assumptions. Thus, there is no profitable deviation at any $p < \hat{p}$.

Case 2: $p + dp \geq \hat{p}$ In that case, if the firm deviates and no breakthrough or exposure occurs, the resulting posterior falls into the continuation region. For that reason, the firm's net present value from deviation is equal to $J(p)$, i.e. the value of never stopping when the belief is p. But we have already argued that $J'(p) > 0$ thus $J(p) < J(\hat{p}) = D$, and this is not a profitable deviation. This concludes the proof of Lemma 36. The firm's value function is:

$$V(p; D) = \begin{cases} J(p), & \text{if } p \geq \hat{p}(D) \\ D, & \text{otherwise} \end{cases}$$

and the state-b continuation value is $V_b = V(p_0; D)$. □

Optimal regulation

To find the regulator's optimal policy, define $\hat{D}$ as the *minimum* value of D such that $\hat{p}(D) = p_0$.

$$\hat{p}(D) = p_0 \iff \hat{D} = p_0(\hat{\alpha} - \hat{\gamma}) + \hat{\gamma} \tag{4.33}$$

This is the smallest value of D that induces immediate disclosure, and so is the value of D that minimizes V_b subject to inducing $T = 0$. The firm's continuation value V_b under this policy is given by:

$$V(p_0; \hat{D}) = J(p_0) \tag{4.34}$$

Any smaller value $D < \hat{D}$ will induce $T = \infty$, but will **not** lower V_b, since the firm can always guarantee itself a payoff of $J(p_0)$. Thus, the regulator has no incentive to offer a value $D < \hat{D}$. This policy achieves $T^* = 0$ and *if* any policy can maintain IC_x, then D^* can maintain it, too. This proves the Proposition.

Appendix 4.D Delay-Dependent Penalties

4.D.1 Preliminary result

Lemma 37. *The two expressions of indifference curves are equivalent, i.e. for* $\Delta T := T' - T$ *(regardless of sign):*

$$U_b(T, D(T), p_0) < U_b(T', D(T'), p_0) \iff$$
$$D(T) < U_b(\Delta T, D(T'), p(T; p_0))$$

Proof. Suppose the agent has belief p_0 at the onset of the breach. For any T', T,

we can write:

$$U_b(T, D(T), p_0) = U_b\Big(T', U_b(T - T', D(T), p(T', p_0)), p_0\Big) \tag{4.35}$$

The left-hand side is the value of waiting until time T. The right-hand side decomposes this into the value of waiting up to time T' and then waiting an additional $\Delta T = T - T'$. At time T', the agent earns the present value of waiting ΔT to receive $D(T)$, given his *current* belief $p(T', p_0)$. *Indifference* between stopping at T and at T' requires:

$$U_b(T, D(T), p_0) = U_b(T', D(T'), p_0) \tag{4.36}$$

Next, equate the right-hand sides of the identity and the indifference condition. The monotonicity of U_b in the second argument yields:

$$D(T') = U_b(T - T', D(T), p(T', p_0)) \tag{4.37}$$

which is the desired expression that holds for every pair of T', T. □

4.D.2 Proof of Lemma 34

By Lemma 37, if the firm is indifferent at the onset of a breach between stopping times t and $t' \leq t$, it must also be indifferent at time t' between stopping or continuing for additional $(t - t')$, given that current beliefs are $p(t'; p_0)$:

$$D(t) = U_b(t - t', D', p(t'; p_0)) \tag{4.38}$$

Recall that the first argument of U_b always is the **time until disclosure**. Observe that (1) for any time t' spent in state b, $p(t'; H) > p(t'; L)$ and that (2) U_b is strictly increasing in p for positive first argument and $D < \alpha$. Thus, if after time tt, type

L prefers to continue for additional $\Delta t = t - t' > 0$, it must be that:

$$D(t) < U_b(t - t', D', p(t'; L)) < U_b(t - t', D', p(t'; H))$$

and type H *also* prefers that. So, that type's stopping time must be weakly higher.

4.D.3 Proof of Lemma 35

To find the indifference curve around point $(t, D(t))$ of type with prior p_0 observe that at time t after the breach occurred, the firm must be indifferent between stopping for $D(t)$ and continuing up to time t', given the *current* belief $p(t; p_0)$.

$$D(t) = U_b(t' - t, D^{p_0}(t'), p(t; p_0))$$

The payment $D^{p_0}(t')$ required to make type $p_0 \in \{L, H\}$ indifferent between stopping at t and continuing up to t' must satisfy the above equation. This expression is algebraically valid for any $t' \in [0, \infty)$. At point $(t, D(t))$, the indifference curves of the two types intersect, hence:[28]

$$U_b(t' - t, D^H(t'), p(t; H)) = U_b(t' - t, D^L(t'), p(t; L))$$

For $t' > t$, U_b is increasing in the current belief. Since $p(t; L) < p(t; H)$, the only way the equality can hold is if $D^H(t') < D^L(t')$. The opposite holds for $t' < t$, hence $D^H(t') < D^L(t')$ if and only if $t' > t$, which concludes the proof.

4.D.4 Proof of Proposition 24

Step 1: Two-Point Contracts.

We can restrict attention to two-point contracts (2PCs) that are incentive compatible (IC). Assume that starting from any schedule $D(t) : [0, \infty) \to R_+$, an

[28] $U_b(0, D(t), p(t; p_0)) = D(t)$, for $p_0 \in \{H, L\}$, since the first argument is time until disclosure.

equilibrium is induced in which type L stops at T_L and receives $D(T_L)$, and type H stops at T_H and receives $D(T_H)$. Replacing this schedule with one that has $D^*(T_L) = D(T_L)$, $D^*(T_H) = D(T_H)$ and $D^*(t) = 0$ every where else, will maintain IR constraints of both types, assuming the original combinations $(T_L, D(T_L))$ and $(T_H, D(T_H))$ respect those constraints. The new schedule $D^*(t)$ is also incentive-compatible: type L prefers stopping at T_L rather than T_H; that was already the case under schedule D and this comparison does not depend on values of D except at $T \in \{T_L, T_H\}$. Type L also prefers stopping at T_L to stopping at any other time. That was already the case under contract D and under D^*, the disclosure payoffs $D(t), t \notin \{T_L, T_H\}$ are weakly lower in D^* than in D. The payoff $U_b(T, D(T), p)$ does not depend on values of D other than $D(T)$, hence each type receives the same interim expected payoff as under the old contract.

Step 2: Binding IC_L, $T_L = 0$

Starting from a given 2PC $\{(T_L, D_L), (T_H, D_H)\}$ that satisfies IC and IR of both types, we can improve the regulator's welfare by replacing the point (T_L, D_L) with $(0, D')$, where $D' = U_b(T_H, D_H, L)$, i.e. constraint IC_L binds. Doing so makes type L indifferent between stopping at $T_L = 0$ and T_H and maintains weakly *lower* interim expected utility. Under the old contract, incentive compatibility requires that $U_b(T_L, D_L, L) \geq U_b(T_H, D_H, L)$ and the right-hand side is the interim expected utility of type L under the new contract. Under the new contract, direct application of Lemma 35 (single-crossing property) guarantees that validity of IC_H is maintained. Importantly, the new contract will also respect IR_L. Since point (T_H, D_H) lies weakly above IR_H:

$$U_b(T_H, D_H, H) \geq (\alpha - \gamma)H + \gamma \iff D_H \geq (\alpha - \gamma)h(T_H) + \gamma$$

The inequality is an application of Lemma 37, applying expression (4.37) as $T \to$

∞ and $\Delta T = T - T_H \to \infty$, too. But we know that $h(T_H) > \ell(T_H)$, hence

$$D(T_H) > (\alpha - \gamma)\ell(T_H) + \gamma$$

By the same argument, this means that point (T_H, D_H) is also above the IR of type L.

$$D_H > (\alpha - \gamma)\ell(T_H) + \gamma \iff U_b(T_H, D_H, L) > (\alpha - \gamma)L + \gamma$$

Type L prefers $U_b(T_H, D_H, L)$ to never disclosing, and by virtue of binding IC_L, also prefers immediate disclosure and D' to never disclosing. Finally, the new contract will also by construction respect limited liability, i.e. $D' \geq 0$. The principal benefits both from a reduction in the stopping time, but also a weak reduction in the firm's continuation payoff $\mathbb{E}V_b$.

Step 3: Binding IR_H

Next, I show that starting from an IC-2PC with $T_L = 0$, the principal can improve welfare if IR_H is slack. IR_H is initially slack iff:

$$U_b(T_H, D_H, H) > U_b(\infty, H) \iff D_H > (\alpha - \gamma)h(T_H) + \gamma$$

by Lemma 37, where $h(T_H) = p(T_H, H)$ is the H type's posterior belief after time T_H of being in state b. The principal can implement a new contract with the same stopping times $T_L = 0$, T_H, but with different payoffs $D'_H = (\alpha - \gamma)h(T_H) + \gamma$ and $D'_L = U_b(T_H, D'_H, L)$, so that IR_H and IC_L bind. Doing so will unambiguously reduce the ex-ante continuation value V_b and increase ex-ante investment incentives. Type H is obviously worse off, and type L is strictly worse off too, since $D'_L = U_b(T_H, D'_H, L) < U_b(T_H, D_H, L) = D_L$; this is true since U_b is always increasing in D, strictly so if $T_H < \infty$.

Again, direct application of Lemma 35 implies that IC_H is maintained under

the new contract. By the same argument as in Step 2, constraint IR_L is again satisfied strictly, since IR_H binds. Limited liability is respected if and only if $D'_L \geq 0$, which is true since $D'_L = U_b(T_H, D_H, L)$ and $D'_H > 0$.

Existence of a trade-off

I prove the following lemma:

Lemma 38. *The total derivative with respect to T_H of $V_b(L; T_H) = U_b(T_H, D(T_H), L)$ is **negative**. As T_H increases in $[0, \infty)$, the interim expected utility of the L type decreases from $(\alpha - \gamma)H + \gamma$ to $(\alpha - \gamma)L + \gamma$.*

Proof. Repeating the equation of L's indifference curve that goes through point (T_H, D_H) – i.e. the equation defining the IC constraint for type L:

$$D(t) = U_b(T_H - t, D(T_H), \ell(t)) \tag{4.39}$$

where $l(t)$ is the L type's posterior after spending time t in state b, with $\ell(0) = L$. In all contracts of Proposition 24, type L receives $D(0) = U_b(T_H, D(T_H), L)$, and $D(T_H) = (\alpha - \gamma)h(T_H) + \gamma$. I show that $D(0)$ is *decreasing* in H's stopping time:

$$\frac{d(D(0))}{dT_H} = \underbrace{\frac{\partial U_b}{\partial T_H}}_{(?)} + (\alpha - \gamma) \underbrace{\frac{\partial U_b}{\partial D} \frac{\partial h(T_H)}{\partial T_H}}_{(-)} \tag{4.40}$$

The first effect is that of prolonging experimentation of type L, holding the disclosure payoff fixed. It is of ambiguous sign. The second effect is that of reducing the payment $D(T_H)$ as we increase T_H and move along the IR curve of type H. This is always negative, since IR_H is downward sloping; after longer time spent in state b, the firm is less optimistic and thus willing to stop for a lower disclosure payoff. Observe that:

1. Evaluated at $L = H$, the above total derivative must be equal to **zero**, since

$U_b(T_H, D(T_H), H) = (\alpha - \gamma)H + \gamma$, independent of T_H. Differentiating again with respect to L yields:

$$\frac{\partial}{\partial L}\left[\frac{dD(0)}{dT_H}\right] = \frac{\partial^2 U_b}{\partial T_H \partial L} + (\alpha - \gamma)\frac{\partial^2 U_b}{\partial D \partial L}\underbrace{\frac{\partial h(T_H)}{\partial T_H}}_{(-)}$$

2. The partial derivative $\frac{\partial^2 U_b}{\partial T_H \partial L}$ is **positive**, for $D \in (\gamma, \alpha)$, higher beliefs increase the marginal benefit to experimentation.

3. The partial derivative $\frac{\partial^2 U_b}{\partial D \partial L}$ is **negative**, for all parameter values. Holding the stopping time fixed, increases in D matter less when beliefs are higher and it becomes more likely that the firm exits before receiving D.

These facts imply that the total derivative in (4.40) is **increasing** in the prior belief of L and always negative for $L < H$. □

Appendix 4.E Contractible investment

Consider the case in which investment is ex-post verifiable and contractible, meaning that the regulator can condition termination payoffs D and E on the level of investment, $D(x) = \{D(0), D(1)\}$. It remains optimal for the regulator to use $E(1) = E(0) = 0$, but it will also be optimal to have $D(1) \geq D(0)$. Allowing the regulator to condition the payoff for voluntary disclosure on the level of investment allows them to provide incentives for early disclosure without jeopardizing ex-ante investment incentives. In fact, increasing the level of $D(1)$ means not only that the stopping time for the firm is earlier following a choice $x = 1$, but also means that the ex-ante incentive to invest is *greater*. I provide a simple proof of the following statement.

Lemma 39. *If the regulator can condition payments D, E on the value of investment, then $E(1) = E(0) = D(0) = 0$ and sufficiently large $D(1)$ always induce*

$x^* = 1$ *and* $T^* = 0$.

Proof. It remains true by usual arguments that $E(1) = E(0) = 0$ is optimal. It also holds that for a firm that chooses $x = 1$, there always exists sufficiently large $D(1)$ that will induce immediate disclosure of a breach. Without loss, the regulator can set $D(0) = 0$, respecting limited liability off-path, too. Define:

$$V_1 := V_b(D(1))$$

$$V_0 := V_b(D(0))$$

as the state b continuation values that firms with different levels of investment face. The firm invests in security iff:

$$-C + \frac{\pi_s + h_1 V_1}{r + h_1} > \frac{\pi_s + h_0 V_0}{r + h_0} \tag{4.41}$$

For sufficiently large $D(1)$, immediate disclosure implies $V_1 = D(1)$, and the above inequality then shows that a sufficiently large $D(1)$ will satisfy IC_x, too. □

Appendix 4.F Derivation of value function under delay-invariant payoffs

A firm that has just been breached and faces delay-invariant payoffs D and E faces an optimal stopping problem in which the state is the current belief that $\theta = 1$. Think of a firm employing a stopping policy with threshold $p^* \in (0, 1)$. Then, the value of that policy for the firm is given by:

$$V(p) = \begin{cases} \tilde{V}(p) & \text{if } p \geq p^* \\ D & \text{if } p < p^* \end{cases}$$

By the Principle of Optimality, the value function at the continuation region, $\tilde{V}$, must satisfy the following Hamilton-Jacobi-Bellman Equation:

$$\tilde{V}(p) = \pi_b dt + e^{-rdt} E(V(p + dp)|p) \tag{4.42}$$

The expectation in the above term is over the uncertain evolution of beliefs. If the firm does not disclose over a time interval dt, he receives flow payoff π_b and termination payoff E with probability μdt, termination payoff S with probability $p\lambda dt$. With complementary probability, and assuming that p is sufficiently above the threshold, he remains in state b, thus the expectation term is given by $E(V(p+dp)|p) = \lambda p dt S + \mu dt E + (1 - \lambda p dt - \mu dt)[\tilde{V}(p + dp)]$. We use a first-order approximation and the derived law of motion for beliefs to get $\tilde{V}(p + dp) = \tilde{V}(p) + \tilde{V}'(p)dp$. Rearranging the Bellman Equation, using the approximation $e^{-rdt} = (1 - rdt)$ and ignoring $o(dt)$ terms yields:

$$\frac{r + \mu + \lambda p}{\lambda p(1 - p)} \tilde{V}(p) + \tilde{V}'(p) = \frac{\pi_b + \lambda p S + \mu E}{\lambda p(1 - p)} \tag{4.43}$$

The above is a non-linear, first-order ordinary differential equation which we need to solve for the value function in the continuation region $p \geq p^*$. We multiply by the integrating factor $I(p) = p^{(\mu+r)/\lambda}(1 - p)^{-(\mu+r+\lambda)/\lambda}$ to get the solution:

$$\tilde{V}(p) = C(1 - p)\omega(p)^{\kappa} + \beta p + \gamma \tag{4.44}$$

where I have defined the odds ratio function:

$$\omega(p) := \frac{1 - p}{p}$$

and the parameters:

$$\beta := \frac{\lambda(\mu+r)S - \lambda(\pi_b + \mu E)}{(\mu+r)((\lambda+\mu+r)} \qquad \gamma := \frac{\pi_b + \mu E}{\mu + r} \qquad \kappa := \frac{\mu + r}{\lambda}$$

To find the constant of integration, C, we use the *smooth pasting*, i.e $\tilde{V}'(p^*) = 0$ and *value matching* $\tilde{V}(p^*) = D$ conditions that the optimal policy must satisfy. Doing that yields the following expressions for the value function and the indifference posterior:

$$p^* = \frac{\kappa(D-\gamma)}{\beta(\kappa+1) - (D-\gamma)} = \frac{(\mu+r)D - (\pi_b + \mu E)}{\lambda(S-D)} = \frac{\kappa(D-\gamma)}{(S-D)} \tag{4.45}$$

$$\tilde{V}(p) = \frac{1-p}{1-p^*}\left(\frac{\omega(p)}{\omega(p^*)}\right)^{\kappa}(D - \gamma - \beta p^*) + \beta p + \gamma \tag{4.46}$$

Looking at the solution for p^*, we can see it lies between $(0, 1)$ if and only if $D \in (\gamma, \beta + \gamma)$. The lower bound is the expected npv for a firm that never discloses and has belief $p = 0$, while $\beta + \gamma$ is the expected npv for a firm that never discloses and has belief $p = 1$.

4.F.1 Slope of value function.

After solving for the value of the integrating constant and indifference posterior, the value function over the *continuation* region $p > p^*$, can be written as:

$$V(p) = \frac{\beta p^*}{(\kappa + p^*)\omega(p^*)^{\kappa}}\omega(p)^{\kappa}(1-p) + \beta p + \gamma$$

and the derivative with respect to p is given by:

$$\frac{dV}{dp} = \frac{-\beta p^*}{\kappa + p^*}\left[\frac{\kappa}{p}\left(\frac{\omega(p)}{\omega(p^*)}\right)^{\kappa} + \left(\frac{\omega(p)}{\omega(p^*)}\right)^{\kappa}\right] + \beta$$

which is larger than zero iff:

$$\left(\frac{\omega(p)}{\omega(p^*)}\right)^{\kappa} \frac{(\kappa+p)/(p)}{(\kappa+p^*)/(p^*)} < 1$$

For $p > p^*$, both ratios are smaller than 1, because $\omega'(p) < 0$. Hence, the slope of the value function is positive in the continuation region, so long as p^* is interior.

Appendix 4.G Binding Monotonicity Constraint

I identify the parameter regions for which the optimal 2PC can be implemented with a schedule $D(t)$ that is *decreasing* in time. In parameter regions such that the non-monotonicity does not arise, if they exist, the outcomes implemented by a contract in the family identified in Proposition 24 can also be implemented by using $D(t) = IC_L$. Remember that type L is indifferent between point $(t, D_L(t))$ and the point $(T_H, D(T_H))$ if:

$$D_L(t) = U_b(\overbrace{T_H - t}^{\text{time until disclosure}}, \overbrace{(\alpha-\gamma)h(T_H)+\gamma}^{D(T_H)}, \quad l(t)) \tag{4.47}$$

where $l(t)$ is the L type's posterior at time t. This is the equation of constraint IC_L, for a given point $(T_H, D(T_H))$. We are interested in the slope of this curve. The first argument of U_b is always the time until disclosure, so that at time t, there is $T_H - t$ left if the agent plans to disclose at T_H. An increase in t has two effects on the indifference payment $D_L(t)$, one via decreasing the time left until disclosure, and one via decreasing the posterior belief $\ell(t)$.

$$\frac{d(D_L(t))}{dt} = \underbrace{-\left.\frac{\partial U_b}{\partial T}\right|_{(T_H-t),l(t)}}_{(?)} + \underbrace{\frac{\partial U_b}{\partial p}\frac{\partial l(t)}{\partial t}}_{(-)} \tag{4.48}$$

My argument is that **(1)** the second component of (4.48) is *zero* at $t = T_H$, **(2)** there is a unique threshold for the prior of the low type, $L^* \in (0, H)$, such that the first component is *positive* at $t = T_H$ iff $L < L^*$, and **(3)** the threshold L^* is independent of T_H; hence, if $L < L^*$, then no 2PC of Proposition 24 is associated with a decreasing IC_L curve.

Step 1: The term $\frac{\partial U_b}{\partial p}(T_H - t,\, D(T_H)\,, \ell(t))$ is zero at $t = T_H$ because time until disclosure is $T_H - t = 0$, thus beliefs do not affect U_b.

Step 2: We can sign the first term at $t = T_H$. Holding $E(t) = 0$ for all t, we already know the ambiguous sign from Proposition 19. At time T_H, i.e. after time $T_H - t$ has passed, the L type would benefit by postponing disclosure by dt if his posterior belief is *greater* than the threshold:

$$\frac{\partial U_b}{\partial T}|_{(T_H - t), l(t)} > 0 \iff \ell(T_H) > \frac{(\mu + r)D(T_H) - \pi_b}{\lambda(S - D(T_H))} \tag{4.49}$$

We know that $h(T_H) \geq \ell(T_H)$ for $L \leq H$. I argue that $h(T_H)$ is greater than the above right-hand side, hence for $L \to H$, the same holds for $\ell(T_H)$. I ask: would the H type prefer to stop at T_H and receive $D(T_H)$, or further hide for dt and then stop to receive $D(T_H)$? By binding IR_H, type H is indifferent between stopping at T_H and **never stopping**. Thus, intuitively, stopping dt later for the same payoff should be strictly preferred (and is easy to confirm formally). By continuity, it must be that if L is sufficiently close to H, type L also **benefits** from prolonging hiding. At the same time, for low enough values of L, the inequality is reversed, because $\ell(T_H)$ is below the threshold at time T_H. Since $\ell(T_H)$ is strictly increasing in the prior, it must be that there is a unique prior $L^* \in [0, H]$, such that $\frac{\partial U_b}{\partial T} > 0$ (at $T = T_H - t$ and $p = l(t)$) if and only if $L > L^*$.

Step 3: Using again the fact that $D(T_H) = (\alpha - \gamma)h(T_H) + \gamma$, we can make further progress in identifying the threshold L^* as a function of H and other parameters.

Evaluating inequality $l(T_H) < \frac{(\mu+r)D_H - \pi_b}{\lambda(S-D_H)}$ we can get the explicit expression for the frontier:

$$L < \frac{H}{1 + (1/\kappa)(1-H)} \tag{4.50}$$

where $\kappa = (\mu + r)/\lambda$.

Importantly, note that the sign of $\frac{\partial U_b}{\partial T}(T_H - t,\, D(T_H)\,, \ell(t))$ does *not* depend on t, as is made clear by (4.49). Hence, if $\frac{\partial U_b}{\partial T}(T_H - t,\, D(T_H)\,, \ell(t)) > 0$ at $t = T_H$ and IC_L is decreasing at T_H, the $D_L(t)$ curve is everywhere decreasing because both components of (4.48) are negative.

Finally, inequality (4.50) does not depend on the stopping time T_H, i.e. on the choice of 2PC. Thus, whether the 2PC outcome can be implemented using a decreasing schedule does not depend on the stopping time T_H that the principal wants to induce.

Putting everything together, we have shown that if $L > L^*$, *any* 2PC outcome can be implemented via a decreasing schedule $D(t)$. Otherwise, this is false for *every* 2PC.

Bibliography

Peter Achim and Jan Knoepfle. Relational enforcement. *Theoretical Economics*, 19(2):823–863, 2024. doi: 10.3982/TE5183. URL https://econtheory.org/ojs/index.php/te/article/view/20240823.

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.

Acronis. The role of cybersecurity insurance in ransomware protection. https://www.acronis.com/en-eu/blog/posts/cybersecurity-insurance-role-in-ransomware-protection/, 2025. Accessed: 2025-07-24.

Toni Ahnert, Michael Brolley, David Cimon, and Ryan Riordan. Cyber risk and security investment. Staff Working Papers 22-32, Bank of Canada, Jul 2022a. URL https://ideas.repec.org/p/bca/bocawp/22-32.html.

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

Eli Amir, Shai Levi, and Tsafrir Livne. Do firms underreport information on cyber-attacks? Evidence from capital markets. *Review of Accounting Studies*, 23(3):1177–1206, 2018. doi: 10.1007/s11142-018-9452-4. URL https://ideas.repec.org/a/spr/reaccs/v23y2018i3d10.1007_s11142-018-9452-4.html.

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.

Terrence August, Duy Dao, Marius Florin Niculescu, and Kihoon Kim. The impact of cryptocurrency on cybersecurity. *Management Science*, 2025. Forthcoming.

Emmanuelle Auriol, Erling Hjelmeng, and Tina Søreide. Corporate criminals in a market context: enforcement and optimal sanctions. *European Journal of Law and Economics*, 56(2):225–287, 2023. doi: 10.1007/s10657-023-09773-w. URL https://ideas.repec.org/a/kap/ejlwec/v56y2023i2d10.1007_s10657-023-09773-w.html.

Christopher Avery and Margaret Meyer. Reputational incentives for biased evaluators. *Unpublished manuscript*, 2012.

Anirudha Balasubramanian. Insurance against ransomware, February 2021. URL https://ssrn.com/abstract=3846111. Available at SSRN: https://ssrn.com/abstract=3846111.

Heski Bar-Isaac and Steven Tadelis. Seller reputation. *Foundations and Trends(R) in Microeconomics*, 4(4):273–351, 2008.

BBC News. Watchdog to fine nhs it firm £6m after medical records hack. BBC News Article, ID c78llg7n5d5o, 2024. URL https://www.bbc.com/news/articles/c78llg7n5d5o. Accessed July 31, 2025; publication date not specified.

Roland Benabou and Guy Laroque. Using privileged information to manipulate markets: Insiders, gurus, and credibility. *The Quarterly Journal of Eco-*

nomics, 107(3):921–958, 1992. URL https://EconPapers.repec.org/RePEc:oup:qjecon:v:107:y:1992:i:3:p:921-958.

Rainer Böhme and Galina Schwartz. Modeling cyber-insurance: Towards a unifying framework. In *Workshop on the Economics of Information Security (WEIS)*, Harvard University, Cambridge, MA, June 2010. Working paper; slides available online.

Patrick Bolton and David S. Scharfstein. A theory of predation based on agency problems in financial contracting. *The American Economic Review*, 80(1):93–106, 1990. URL http://www.jstor.org/stable/2006736.

Oliver Bullough. Sewage sleuths: the men who revealed the slow, dirty death of welsh and english rivers. The Guardian– Environment section, August 2022. URL https://www.theguardian.com/environment/2022/aug/04/sewage-sleuths-river-pollution-slow-dirty-death-of-welsh-and-english-rivers. Published August 4, 2022 UTC; accessed July 31, 2025.

Tracey Caldwell. Can you put a dollar amount on your company's cyber risk? *Harvard Business Review*, October 19 2016. URL https://hbr.org/2016/10/can-you-put-a-dollar-amount-on-your-companys-cyber-risk.

Anna Cartwright and Edward Cartwright. Ransomware and reputation. *Games*, 10(2):1–14, June 2019. URL https://ideas.repec.org/a/gam/jgames/v10y2019i2p26-d238459.html.

Anna Cartwright, Edward Cartwright, Jamie MacColl, Gareth Mott, Sarah Turner, James Sullivan, and Jason R.C. Nurse. How cyber insurance influences the ransomware payment decision: Theory and evidence. *The Geneva Papers on Risk and Insurance–Issues and Practice*, 48(2):300–331, 2023. doi: 10.1057/s41288-023-00288-8.

Computer Weekly. Cover-ups still the norm in the wake of a cyber incident. Computer Weekly, September 2023. URL https://www.computerweekly.com/news/366553240/Cover-ups-still-the-norm-in-the-wake-of-a-cyber-incident. Accessed July 31, 2025.

Will Cong, Campbell Harvey, Daniel Rabetti, and Zong-Yu Wu. An anatomy of crypto-enabled cybercrimes. *Management Science*, 71(4):3622–3633, 2025. doi: 10.1287/mnsc.2023.03691. URL https://doi.org/10.1287/mnsc.2023.03691.

Matteo Crosignani, Marco Macchiavelli, and Andre F. Silva. Pirates without borders: The propagation of cyberattacks through firms' supply chains. *Journal of Financial Economics*, 147(2):432–448, 2023. doi: 10.1016/j.jfineco.2022.12.002. URL https://ideas.repec.org/a/eee/jfinec/v147y2023i2p432-448.html.

Cybersecurity and Infrastructure Security Agency. Information sharing. CISA website. URL https://www.cisa.gov/topics/cyber-threats-and-advisories/information-sharing. Accessed July 31, 2025.

Alexandre de Cornière and Greg Taylor. A Model of Information Security and Competition. *Marketing Science*, forthcoming, 2024. URL https://pubsonline.informs.org/doi/10.1287/mksc.2023.0513.

Miguel Godinho de Matos and Idris Adjerid. Consumer consent and firm targeting after gdpr: The case of a large telecom provider. *Management Science*, 68(5):3330–3378, 2022. URL https://EconPapers.repec.org/RePEc:inm:ormnsc:v:68:y:2022:i:5:p:3330-3378.

Douglas Diamond. Reputation acquisition in debt markets. *Journal of Political*

Economy, 97(4):828–62, 1989. URL https://EconPapers.repec.org/RePEc:ucp:jpolec:v:97:y:1989:i:4:p:828-62.

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.

Ronald A. Dye. Disclosure of non-proprietary information. *Journal of Accounting Research*, 23(1):123–145, 1985. URL https://EconPapers.repec.org/RePEc:bla:joares:v:23:y:1985:i:1:p:123-145.

Euronews. Cybercrime: Insurance giant axa to stop covering ransomware payments in france, May 2021. URL https://www.euronews.com/2021/05/07/cybercrime-insurance-giant-axa-to-stop-covering-ransomware-payments-in-france. Accessed: 2025-07-24.

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.

Federal Trade Commission. Ftc takes action against marriott and starwood over multiple data breaches. FTC News Release, October 2024. URL https://www.ftc.gov/news-events/news/press-releases/2024/10/ftc-takes-action-against-marriott-starwood-over-multiple-data-breaches. Accessed July 31, 2025.

Financial Times. The ransomware battle is shifting — so should our response, 2024. URL https://www.ft.com/content/3b172a2a-4be5-4ef4-87cb-7fdcdee2ad99. Opinion piece on strategic shifts in ransomware defense.

Financial Times. M&s cyber insurance payout to be worth up to £100mn, 2025. URL https://www.ft.com/content/

723b6195-1ce7-4b5f-94f5-729e9152c578. Coverage of Marks & Spencer's potential £100m cyber-insurance claim and implications for premiums.

Jens Foerderer and Sebastian W. Schuetz. Data Breach Announcements and Stock Market Reactions: A Matter of Timing? *Management Science*, 68(10):7298–7322, 2022. doi: 10.1287/mnsc.2021.4264. URL https://ideas.repec.org/a/inm/ormnsc/v68y2022i10p7298-7322.html.

Neil Gandal, Michael Riordan, and Shalom Bublil. A New Approach to Quantifying, Reducing and Insuring Cyber Risk: Preliminary Analysis and Proposal for Further Research. CEPR Discussion Papers 14461, C.E.P.R. Discussion Papers, March 2020. URL https://ideas.repec.org/p/cpr/ceprdp/14461.html.

Neil Gandal, Tyler Moore, Michael Riordan, and Noa Barnir. Empirically Evaluating the Effect of Security Precautions on Cyber Incidents. CEPR Discussion Papers 17605, C.E.P.R. Discussion Papers, October 2022. URL https://ideas.repec.org/p/cpr/ceprdp/17605.html.

GDPR.eu. What are the gdpr consent requirements? GDPR.eu Article, May 2018. URL https://gdpr.eu/gdpr-consent-requirements/. Accessed July 31, 2025.

Tom Gerken. Ex-uber security chief sentenced over covering up hack. BBC News – Technology section, May 2023. URL https://www.bbc.com/news/technology-65497186. Published May 5, 2023; accessed July 31, 2025.

Avi Goldfarb and Catherine Tucker. Introduction to The Economics of Privacy . *National Bureau of Economic Research Conference Report*, May 2023. URL https://ideas.repec.org/h/nbr/nberch/14786.html.

Matthew Gooding. Is ransomware insurance fuelling the ransomware boom?, jul 2024. URL https://www.techmonitor.ai/technology/ransomware-insurance?cf-view.

Marina Halac, Navin Kartik, and Qingmin Liu. Optimal Contracts for Experimentation. *The Review of Economic Studies*, 83(3):1040–1091, 2016. URL https://ideas.repec.org/a/oup/restud/v83y2016i3p1040-1091.html.

Seth Hastings, Tyler Moore, Neil Gandal, and Noa Barnir. Measuring user costs of enterprise multifactor authentication policies. *Available at SSRN*, 2023. URL http://dx.doi.org/10.2139/ssrn.4669442.

Daniel N Hauser. Censorship and reputation. *American Economic Journal: Microeconomics*, 15(1):497–528, 2023.

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

Shota Ichihashi. Dynamic Privacy Choices. *American Economic Journal: Microeconomics*, 15(2):1–40, May 2023. doi: 10.1257/mic.20210100. URL https://ideas.repec.org/a/aea/aejmic/v15y2023i2p1-40.html.

Roman Inderst and Holger M. Mueller. CEO replacement under private information. *Review of Financial Studies*, 23(8):2935–2969, 2010. URL https://EconPapers.repec.org/RePEc:oup:rfinst:v:23:y:2010:i:8:p:2935-2969.

Information Commissioner's Office. Data security incident trends. Information Commissioner's Office – Complaints and Concerns Data Sets and Trends Reports, May 2025a. URL https://ico.org.uk/action-weve-taken/complaints-and-concerns-data-sets/data-security-incident-trends/. Data updated to Q1 2025 (latest report released May 6, 2025); accessed July 31, 2025.

Information Commissioner's Office. What is valid consent? Information Commissioner's Office website, June 2025b. URL https:

//ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/lawful-basis/consent/what-is-valid-consent/. Accessed July 31, 2025.

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 target firms. *Journal of Financial Economics*, 139(3):719–749, 2021a.

Shinichi Kamiya, Jun-Koo Kang, Jungmin Kim, Andreas Milidonis, and René M. Stulz. Risk management, firm reputation, and the impact of successful cyberattacks on target firms. *Journal of Financial Economics*, 139(3):719–749, 2021b. URL https://EconPapers.repec.org/RePEc:eee:jfinec:v:139:y:2021:i:3:p:719-749.

Louis Kaplow and Steven Shavell. Optimal law enforcement with self-reporting of behavior. *Journal of Political Economy*, 102(3):583–606, 1994. URL https://EconPapers.repec.org/RePEc:ucp:jpolec:v:102:y:1994:i:3:p:583-606.

Sam Kapon. Dynamic amnesty programs. *American Economic Review*, 112(12): 4041–75, 2022.

Anil K. Kashyap and Anne Wetherilt. Some principles for regulating cyber risk. *AEA Papers and Proceedings*, 109:482–487, 2019. URL https://ideas.repec.org/a/aea/apandp/v109y2019p482-87.html.

Michael L. Katz. Game-playing agents: Unobservable contracts as precommitments. Economics Working Papers 91-172, University of California at Berkeley, Jul 1991. URL https://ideas.repec.org/p/ucb/calbwp/91-172.html.

Godfrey Keller, Sven Rady, and Martin Cripps. Strategic experimentation with exponential bandits. *Econometrica*, 73(1):39–68, 2005. URL https://EconPapers.repec.org/RePEc:ecm:emetrp:v:73:y:2005:i:1:p:39-68.

Aniket Kesari. Do data breach notification laws work? *New York University Journal of Legislation and Public Policy*, 26:173–228, 2024.

Sang-Hyun Kim. Time to Come Clean? Disclosure and Inspection Policies for Green Production. *Operations Research*, 63(1):1–20, 2015. doi: 10.1287/opre.2015.1345. URL https://ideas.repec.org/a/inm/oropre/v63y2015i1p1-20.html.

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.

David Kreps and Robert Wilson. Reputation and imperfect information. *Journal of Economic Theory*, 27(2):253–279, 1982.

Jean-Jacques Laffont and Jean Tirole. *A Theory of Incentives in Procurement and Regulation*. MIT Press, Cambridge, MA, 1993. ISBN 978-0262121743. URL https://mitpress.mit.edu/9780262121743/a-theory-of-incentives-in-procurement-and-regulation/.

Aron Laszka, Samaneh Farhang, and Jens Grossklags. On the economics of ransomware. In *Decision and Game Theory for Security: 8th International Conference, GameSec 2017, Proceedings*, Vienna, Austria, 2017. doi: 10.1007/978-3-319-68711-7_20.

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.

Steven Levitt and Christopher Snyder. Is no news bad news? information transmission and the role of "early warning" in the principal-agent model. *RAND Journal of Economics*, 28(4):641–661, 1997. URL https://EconPapers.repec.org/RePEc:rje:randje:v:28:y:1997:i:winter:p:641-661.

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

George Mailath and Larry Samuelson. Who wants a good reputation? *The Review of Economic Studies*, 68(2):415–441, 2001. URL https://EconPapers.repec.org/RePEc:oup:restud:v:68:y:2001:i:2:p:415-441.

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.

Marsh. Ransomware, 2025. URL https://www.marsh.com/en/services/cyber-risk/expertise/ransomware.html. Insights into Marsh's ransomware readiness and cyber-insurance expertise.

Tom Meurs, Edward Cartwright, Anna Cartwright, Marianne Junger, Raphael Hoheisel, Erik Tews, and Abhishta Abhishta. Ransomware economics: A two-step approach to model ransom paid. In *2023 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–13, 2023. doi: 10.1109/eCrime61234.2023.10485506.

Amalia Miller and Catherine Tucker. Privacy protection, personalized medicine, and genetic testing. *Management Science*, 64(10):4648–4668, 2018. URL https://EconPapers.repec.org/RePEc:inm:ormnsc:v:64:y:2018:i:10:p:4648-4668.

MIT Technology Review. Facebook data leak. MIT Technology Review, April 7, 2021, April 2021. URL https://www.technologyreview.com/2021/04/07/1021892/facebook-data-leak/. Accessed July 31, 2025.

Munich Re. Cyber insurance: Risks and trends 2025, 2025. URL https://www.munichre.com/en/insights/cyber/cyber-insurance-risks-and-trends-2025.html#379179748. Accessed: 2025-07-24.

Michael Mussa and Sherwin Rosen. Monopoly and product quality. *Journal of Economic Theory*, 18(2):301–317, 1978. URL https://EconPapers.repec.org/RePEc:eee:jetheo:v:18:y:1978:i:2:p:301-317.

National Cyber Security Centre. Global ransomware threat expected to rise with ai, may 2024. URL https://www.ncsc.gov.uk/news/global-ransomware-threat-expected-to-rise-with-ai.

National Cyber Security Centre (NCSC). Cyber security advice for large organisations. NCSC website. URL https://www.ncsc.gov.uk/section/advice-guidance/large-organisations. Accessed July 31, 2025.

NPR. After data breach exposes 530 million, facebook says it will not notify users. NPR News, April 2021. URL https://www.npr.org/2021/04/09/986005820/after-data-breach-exposes-530-million-facebook-says-it-will-not-notify-users. Accessed July 31, 2025.

Greg Palmer. Tightening cybersecurity laws won't stop hackers. *The Times*, April 2025. URL https://www.thetimes.com/uk/law/article/tightening-cybersecurity-laws-wont-stop-hackers-lrr3b5g37.

Gregory Parks. Sec's new data breach requirement increases obligations for financial services companies. Morgan Lewis Law Firm website, June 2024. URL https://www.morganlewis.com/pubs/2024/06/

secs-new-data-breach-requirement-increases-obligations-for-financial-services-co Accessed 2025-07-31.

Parra and Ralph A. Winter. Optimal insurance contracts under moral hazard, December 2025. URL https://ideas.repec.org/h/spr/sprchp/978-3-031-69674-9_6.html.

Christian Peukert, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer. Regulatory Spillovers and Data Governance: Evidence from the GDPR. *Marketing Science*, 41(4):746–768, July 2022. doi: 10.1287/mksc.2021.1339. URL https://ideas.repec.org/a/inm/ormksc/v41y2022i4p746-768.html.

A. Mitchell Polinsky and Steven Shavell. The economic theory of public enforcement of law. *Journal of Economic Literature*, 38(1):45–76, March 2000. doi: 10.1257/jel.38.1.45. URL https://www.aeaweb.org/articles?id=10.1257/jel.38.1.45.

Caleb Rawson, Brady J. Twedt, and Jessica C. Watkins. Managers' strategic use of concurrent disclosure: Evidence from 8-k filings and press releases. *The Accounting Review*, 98(4):345–371, 2023. URL https://doi.org/10.2308/TAR-2021-0088.

SentinelOne. What is information sharing in cybersecurity? SentinelOne Cybersecurity 101, 2025. URL https://www.sentinelone.com/cybersecurity-101/cybersecurity/what-is-information-sharing/. Accessed July 31, 2025;.

A. Spence. Monopoly, quality, and regulation. *Bell Journal of Economics*, 6(2):417–429, 1975. URL https://EconPapers.repec.org/RePEc:rje:bellje:v:6:y:1975:i:autumn:p:417-429.

The Guardian. M&s says some personal data was taken in cyber-attack, 2025. URL https://www.theguardian.com/business/2025/may/13/

m-and-s-personal-data-cyber-attack-marks-spencer-card-passwords. Report on Marks & Spencer data breach: customer names, addresses and order histories exposed; card details and passwords unaffected.

Ying Lei Toh. Incentivizing firms to protect consumer data: Can reputation play a (bigger) role? *PhD Thesis, Toulouse School of Economics*, 2018.

G-J van Rooyen. What is double extortion ransomware?, jun 2024. URL https://www.techmonitor.ai/technology/cybersecurity/double-extortion-ransomware?cf-view.

Shouqiang Wang, Peng Sun, and Francis de Véricourt. Inducing environmental disclosures: A dynamic mechanism design approach. *Operations Research*, 64 (2):371–389, 2016. doi: 10.1287/opre.2016.1476.

Woodruff Sawyer. The new hacker playbook: Weaponizing the sec's cyber disclosure rules. Online, 2024. URL https://woodruffsawyer.com/insights/sec-cyber-disclosure-rules. Accessed: 2025-08-14.