BANK OF CANADA
BANQUE DU CANADA

Staff Working Paper/Document de travail du personnel—2022-32

Last updated: February 16, 2024

Cyber Risk and Security Investment

by Toni Ahnert,[1] Michael Brolley,[2] David Cimon[3] and Ryan Riordan[4]

[1]European Central Bank
Center for Economic and Policy Research
Finance Theory Group
toni.ahnert@ecb.europa.eu

[2]Lazaridis School of Business and Economics,
Wilfrid Laurier University
mbrolley@wlu.ca

[3]Financial Markets Department
Bank of Canada
DCimon@bank-banque-canada.ca

[4]LMU Munich
ryan.riordan@lmu.de

DOI: https://doi.org/10.34989/swp-2022-32 | **ISSN** 1701-9397

Acknowledgements

This paper supersedes a previous version titled “Cybersecurity and Ransomware in Financial Markets.” We thank Jason Allen, Kartik Anand, Shota Ichihashi, Marco Macchiavelli and Sophie Moinas for thoughtful comments. We thank conference participants at the Northern Finance Association and seminar participants at Wilfrid Laurier University, the University of Toronto, the University of Hawaii, the Bank of Canada and the CFTC, as well as the staff from the Bank of Canada's Resolution and Crisis Preparedness team for their helpful feedback. We thank Michael Beckenhauer and William Wootton for excellent research assistance. Michael Brolley acknowledges financial support from the Social Sciences and Humanities Research Council, Grant No. 430-2019-00814.

Abstract

We develop a model in which firms invest in cybersecurity to protect themselves and their clients from cyber attacks. Since cyber security investment is unobservable, firms may signal their investment to attract clients. In equilibrium, firms under-invest in cyber security. We derive testable implications for the modality of cyber attacks, the probability of a successful attack, and client fees. To improve efficiency, a regulator can impose a minimum level of security investment or legislate consumer protection that shifts the burden of cyber attacks from clients to firms. Both regulations induce firms to invest the constrained-efficient amount in cyber security.

Topics: Economic Models; Financial System Regulation and Policies; Financial Services; Financial Stability; Payment Clearing and Settlement Systems
JEL codes: D78, D81, G18, G21, G23

Résumé

Nous élaborons un modèle où des entreprises investissent en cybersécurité pour se protéger et protéger leurs clients contre les cyberattaques. Puisque les investissements en cybersécurité ne sont pas observables, les entreprises peuvent signaler leur niveau d'investissement en cybersécurité pour attirer des clients. En situation d'équilibre, les entreprises sous-investissent en cybersécurité. Nous déduisons des implications vérifiables au sujet des modalités des cyberattaques, de la probabilité d'une cyberattaque réussie et des frais imposés aux clients. Pour améliorer l'efficience, une autorité de réglementation peut imposer un niveau minimal d'investissement en cybersécurité ou faire adopter un règlement en matière de protection des consommateurs qui fait porter aux entreprises, et non aux clients, le fardeau des cyberattaques. Dans les deux cas, les entreprises doivent investir un montant sous contrainte d'efficience en cybersécurité.

Sujets : Modèles économiques; Réglementation et politiques relatives au système financier; Services financiers; Stabilité financière; Systèmes de compensation et de règlement des paiements
Codes JEL : D78, D81, G18, G21, G23

1 Introduction

The digital age has improved connectivity between firms and clients but has also introduced new security threats to firms handling client assets and data. Cyber attacks on these firms are frequent and their prevention is expensive. For example, JP Morgan Chase invests $15 billion in cyber security annually to fight off an estimated 45 billion hack attempts per day (CNN 2024). Cyber attacks can have staggering consequences. In December 2019, a breach of Bitmart, a cryptocurrency exchange, resulted in the theft of over $150 million in client assets (Bloomberg 2021). In November 2023, a ransomware attack on ICBC, one of the world's largest banks, disrupted trading in U.S. Treasuries until ransom demands were met (Reuters 2023).[1] Recognizing the substantial costs of security and security vulnerabilities, cyber risk management is a critical competitive factor for firms.

To defend client assets and data against cyber attacks, firms can invest in cyber security. But how do firms convince clients that cyber security investments translate into more secure assets and data? Since cyber security strategies are complex and security protocols are private, clients often lack the sophistication to observe or verify the cyber security investments of firms. How does this impact clients' willingness to pay for security, and thus firms' incentive to invest in it? This issue is top-of-mind at the SEC, given their proposal to enhance public disclosure to "strengthen investors' ability to evaluate public companies' cyber security practices and incident reporting" (Securities and Exchange Commission 2022).

In this paper, we propose a model of cyber security. The unobservability of firm security investment for clients increases firm vulnerability to cyber attacks and lowers welfare. Enhancing the transparency of security investment for clients through signalling technologies, such as cyber security ratings, reduces firm vulnerability and increases welfare. A regulator who sets minimum investment requirements or mandates firm liability for client losses

[1]The FinCyber Project documents over 150 significant cyber incidents at global financial institutions since 2019 (Carnegie 2022). To assess the cyber resilience of their financial system, the European Central Bank currently stress tests the responses of 109 banks for disruption-type attacks (ECB 2024).

associated with cyber attacks attains the constrained-efficient outcome, that is the level of welfare in the benchmark case in which security investment is observable to clients.

In our model described in Section 2, firms facilitate client transactions that are vulnerable to cyber attacks. A cyber attacker has two choices: it first chooses the modality of attack, defining the proportions with which the firms or clients bear the direct costs of the attack. This captures the array of attack methods, from more conventional attacks where attackers steal the clients assets or data to one in which a ransomware attack may disrupt the firm's ability to conduct business. Second, the attacker chooses the intensity of the attack at a cost, and reaps a proportion of the transaction value upon a successful attack.

Clients allocate their transaction needs across firms according to their transaction value-at-risk, which we define as the part of a transaction vulnerable to a cyber attack. The value-at-risk of a transaction does not need be the entire amount of an account or it could represent the opportunity costs of being unable to access an account or complete a transaction.[2] A principal-agent problem arises because clients cannot observe the level of security investment. Firms compete in security investment and the associated fees charged to clients. The probability of a successful cyber attack increases in attack intensity but decreases in security investment, as in classical attacker-defender games (e.g., Goyal and Vigier 2014).

Our model yields stark differences depending on whether security investment is observable. When clients cannot observe investment (Section 3.1), the attacker chooses an attack modality that targets clients directly, anticipating that firms lack the incentives to protect clients by investing in security as they cannot convince clients that they have done so. Hence, vulnerability to attacks is high and welfare is low. By contrast, observable investment (Section 3.2) drastically improves both security investment and welfare—achieving the second-best outcome—as clients are willing to pay for security that they observe. Even in this setting, cyber attacks sometimes succeed, so zero vulnerability to cyber risk is inefficient.

[2]For example, bank accounts have maximum daily transfer amounts limiting the losses to a fraction of the total account value. Similarly, the ICBC hack limited the ability of their clients, most notably BNY Mellon, to trade because their capital was inaccessible for multiple days.

In practice, there is significant unobservability of cyber security investment—consistent with the SEC's stated desire to improve the transparency of investment in cyber security. Hence, we proceed by examining whether market-based or regulatory solutions can improve welfare and achieve the second-best outcome. We start by studying whether firms can pay to credibly signal their level of cyber security investment. In practice, a signal is available to firms through security ratings offered by third-party firms such as BitSight and UpGuard.[3]

With a market-based solution for firms to acquire costly signals of security investment (Section 4), we find that the attacker no longer focuses their attack entirely on clients. Doing so incentivizes firms to protect clients by investing in security and signalling their investment. Instead, the attacker balances their attacks to impact both firms and clients, such that firms have sufficient incentive to invest in security and credibly protect themselves without having to signal. That is, the cyber attacker uses his choice of attack modality to affect the firm's choice of whether to signal its security investment to clients. The availability of a costly signal lowers firm vulnerability and improves welfare but does not reach the second best.

Section 5 examines whether regulation could further improve welfare above the market-based signalling solution. First, we propose a policy that targets cyber attack prevention by influencing security investment directly. In particular, we consider a regulator that mandates a (publicly-observed) minimum standard for cyber security investment. By setting the minimum standard at the level that a firm would choose if clients were to directly observe security investment, the resulting regulated equilibrium achieves the second-best outcome.

Alternatively, a regulator can influence cyber attack prevention by imposing consequences on firms who suffer breaches in which clients face losses, which we label consumer protection regulation. We specify client losses instead of all losses because our model argues that firms are incentivized to protect *themselves* from losses, even when security investment is

[3]BitSight and UpGuard are two U.S.-based firms that provide third-party cyber risk management services to firms and institutions. Both companies offer an assessment of a firm's vulnerability to cyber attacks, which is summarized by a numerical "security score." These systems are similar in design to a credit ratings score. Signal-type information like security ratings are available for both firms and clients to purchase, while firms can purchase more in-depth measures like cyber vulnerability "Value-at-Risk" and breach probability.

unobservable to clients. Suppose a regulator can shift the liability of successful attacks from clients to firms, or who penalizes firms for breaches that impact consumers. Such regulation alters the security investment incentive for firms, and thus the attack modality used by the attacker in equilibrium. When a regulator assigns all of the liability from losses to firms, firms choose to invest in security and the economy reaches the second-best level of welfare. Consumer protection in our model broadly reflects the current legal arrangement in some jurisdictions, such as the European Union's General Data Protection Regulation (GDPR) that allows for the imposition of fines on firms whose security lapses impact consumers.

Finally, our model generates several testable implications. First, our model implies that both direct and indirect costs of firm security provision correlate with the types of attacks observed and subsequent fees paid by clients. When firms face higher security costs, firms are more reluctant to protect clients from breaches, and thus attackers go after client assets and data directly rather than firm-directed service-disruption ransom attacks. Second, it implies that the cost to firms to signal security investment play an important, and perhaps counter-intuitive, role in cyber security. Since attackers do not want firms to publicize their security investment to clients (via costly signals), they reduce their client-focused attacks, so as to reduce the benefit to firms from advertising their protection efforts. An increase in the cost of signalling incentivizes attackers to increase client-focused attacks, as firms are more reluctant to pay the signalling cost. Moreover, the refocus to attacking clients versus firms directly reduces the firms incentive for security provision, leading these savings to be passed through into lower fees, but at a cost of more breaches on average.

Literature. Our paper is related to several strands of literature. First, we contribute to a nascent, but rapidly growing literature on cyber risk. Recent empirical literature focuses primarily on two broad questions: the impact of cyber risk on firm value and stock returns, and the spillover ramifications of cyber attacks. Several papers provide evidence that cyber risk has a detrimental impact on firm valuation and equity returns (e.g., Jamilov et al. (2022),

Akey et al. (2021), Berkman et al. (2018), Garg (2020)). Florackis et al. (2023) describe cyber risk at the firm level and show that increased cyber risk is associated with higher equity returns, which suggests that firms or industries more susceptible to cyber attacks have higher costs of capital to compensate for the additional risk. Kamiya et al. (2021) document that following a successful cyber attack, the decrease in shareholder wealth exceeds that of the out-of-pocket costs of the attack.

Beyond individual firm effects, spillover effects are a primary concern. Duffie and Younger (2019) explore the repercussions of a cyber run on 12 of the largest U.S. financial institutions. Eisenbach et al. (2022) sheds light on the spillover effect of cyber attacks on the U.S. financial system, where an attack on one of the major banks in the country would negatively impact almost a third of the wholesale payment network between U.S. financial institutions. Crosignani et al. (2023) document the contagion effects of a large-scale attack, showing that the costs of such an attack reach far beyond the targeted firm. Kotidis and Schreft (2022) highlight the importance of bank contingency planning towards mitigating spillover effects when attacks continue over a period of days. Kopp et al. (2017) argue, however, that the private market may under-invest in cybersecurity relative to the social optimum, creating room for policy intervention. Our paper contributes a theoretical framework that may inform future empirical work on the joint relationship between cyber risk and security, and provides direction for regulatory interventions to bolster firm investment in cyber security.

The majority of theory work on the economics of cyber risk and cyber security is found in the computer science literature, where several papers highlight the economic incentives of cyber security (see e.g., Anderson (2001); Anderson and Moore (2006); Anderson et al. (2013)). They show, for example, that when banks are liable for losses, more security is provided. In the economics literature, Anand et al. (2022) focuses on systemic risk considerations, such as the ability to cause bank runs, not present in other types of crime. Other areas of computer science and information technology study the security investment problem in different contexts: as a user's responsibility (August and Tunca 2006); as a profit

maximization problem (Dynes et al. 2007); as a function of the importance of a vulnerability (Gordon and Loeb 2002); and in the presence of state actors with almost infinite resources (Anderson 2001). Our contribution is to examine the principal-agent problem inherent in the provision of security by firms that manage client assets, transactions, and data.

Our work also belongs to the class of attacker-defender games in economics. Vásquez (2022) examines the incentives of criminals more broadly, modelling the costly security decisions by potential victims as they attempt to guard against and discourage attackers, who fear punishment. The model recommends that governments focus on mandating increased security by defenders versus increasing penalties against attackers. Our paper is focused on cyber crime and models firms who protect both themselves as well as their clients, while criminals are not dissuaded by the prospect of punishment. We also show that security mandates, among other firm-centric policy options, improve welfare.

Much of the remaining theoretical attacker-defender literature focuses on the structure of networks. For example, Dziubiński and Goyal (2013), Goyal and Vigier (2014), Acemoglu et al. (2016), and Hoyer and de Jaegher (2016) analyze the incentives for attackers and defenders who expend resources to secure nodes and the entire network. We depart from a network structure setup and contagion issues to focus on a single point of failure. This simplification allows us to tractably study the principal-agent problem inherent in many applications of cyber risk and cyber security, which is our main contribution to this literature.

2 Model

There are four dates $t = 0, 1, 2, 3$, no discounting, a single divisible good for consumption and investment, and three types of risk-neutral agents: clients, firms, and an attacker. At $t = 0$ the attacker chooses the attack modality. At $t = 1$ firms invest in cyber security and charge a fee to clients who transact with them. At $t = 2$ clients allocate their transactions across firms. At $t = 3$ the attacker chooses the intensity of the attack of each firm.

When clients transact with a firm, some transaction value V is vulnerable to a cyber attack, which we refer to as (transaction) value-at-risk.[4] In the case where cryptocurrency is transferred from the client to an exchange—as in the 2021 BitMart theft—the total value may be at risk to a cyber attacker who steals the assets under exchange custody. Our notion of value-at-risk also captures that not all of an account or transaction needs to be lost upon a successful hack. Most financial institutions limit the amount of cash that can be transferred per day. In the case of a hacked account, this limits the daily losses to a portion of the account total rather than the total available in the account. In another setting, when ICBC was hacked they were unable to clear and settle client trades in Treasuries, effectively limiting client access to cash and collateral. The inability to access capital impaired clients' ability trade for multiple days potentially leading to losses and forgone gains.[5]

Attacker. The attacker attacks firms to steal the value of transactions at risk to a cyber attack. The risk may be to the clients, firms, or a combination of both. To differentiate from conventional crime[6], the cyber attacker does not fear being caught or punished. This assumption maps to real-world cyber attackers who are difficult to either identify or prosecute based on jurisdictional boundaries when compared to conventional criminals.

At $t = 0$, the attacker chooses the attack modality ℓ_i. The attack modality does not enter the attacker's payoff directly; instead, it determines the allocation of losses between the clients and the firms. If an attack on firm i is successful, clients lose a fraction ℓ_i of the transaction value at risk, while the firm loses a fraction $1 - \ell_i$. Initially, we focus on an attacker with complete control of the division of losses. Later we will also discuss regulatory or legal frameworks that ensure that clients are protected from losses.

[4] A summary of notation used can be found in Appendix A.

[5] The ICBC hack led to a disruption in Treasury markets that saw ICBC sending USB sticks by courier to clear and settle client trades. The hack had a large impact on BNY Mellon that was owed more than $9 billion at one point by ICBC, more than NBY Mellon's total net assets (Reuters 2023)

[6] The classical reference for the economics of crime and punishment is Becker (1968).

We envision two interpretations for the choice of attack modality. First, it may represent the degree of vulnerability that the firm or its clients bear in a successful attack using different methods of attack. As in the 2021 BitMart example, an attacker that successfully steals client assets presents a vulnerability to the firm's clients, but not necessarily the firm itself, representing $\ell_i = 1$. Alternatively, as in the 2023 ICBC attack, a ransomware attack that halts the firm's operations may primarily impact the firm, representing $\ell_i = 0$.

Second, ℓ_i could represent the choice of target. An attacker who chooses a low value of ℓ_i could be seen as choosing to attack firms who incur losses regardless of the type of attack. For example, consumer protection legislation may require firms to make clients whole, or may be required to report breaches and pay fines as a result.[7] In these cases, even if attackers target the firms' clients, the firms still suffer losses. Alternatively, a high value of ℓ_i could represent firms that are unlikely to reimburse clients for losses or damage from cyber incidents.

At $t = 3$, the attacker chooses the attack intensity $a_i \geq 0$ to maximize their payoff

$$\pi_A = \sum_{i=1}^{N} \left(r\delta_i V_i - a_i \right), \tag{1}$$

where V_i is the value-at-risk at firm i, δ_i is the probability of a successful attack on firm i, and $r \in (0, 1]$ is the portion of the value-at-risk that the attacker can steal (their payoff or reward).[8]

We offer two interpretations of r. First, r could be the inherent ease of stealing the asset or a recovery rate. For example, records of physical asset ownership may have $r = 0$, as even if the records are stolen or corrupted, backup copies exist that prevent the transfer. Digital assets (e.g., crypto wallet addresses and banking information) on centralized systems may have a higher r, by contrast, as digital records of asset ownership may be accessed and transactions authorized and cleared before the attacker can be intercepted. The 2021 attack

[7] For example, British Airways was fined £20m for a 2018 data breach under the EU GDPR (FT 2020).

[8] While the reward r is exogenous, both the transaction value-at-risk V_i and the probability of a successful attack δ_i are endogenous in our model.

on Bitmart, and other similar incidents at crypto-currency exchanges, are prime examples of realized attacks at venues where r may be high and digital assets can be stolen.

Second, we can interpret r as the relative value of data or assets that can be stolen or the ease with which they can be monetized by the attacker. This interpretation reflects the disparity between the total value-at-risk to the client and the value of the attack proceeds to the attacker. For example, personal data may offer an attacker the *possibility* of stealing all of a client's assets at risk, but in practice the attacker may not be able to realize the full value of the data before the firm recognizes the breach and denies access.[9]

Firms. At $t = 1$, $N \geq 2$ firms indexed by $i = 1, ..., N$ each simultaneously invest $s_i \geq 0$ in the cyber security of their firm (e.g., hiring an information systems analyst, implementing biometric identification and/or multi-factor authentication) and choose a fee $f_i \geq 0$ per unit of transaction by the firm. Each firm maximizes its expected profits

$$\pi_i = f_i V_i - (1 - \ell_i)\delta_i V_i - cs_i, \tag{2}$$

where c is the cost of security investment. If the firm is successfully attacked, we refer to it as having been *breached*.

Following Goyal and Vigier (2014), an attack on firm i is successful with probability

$$\delta_i = \delta(a_i, s_i) = \frac{a_i}{a_i + s_i} \tag{3}$$

if $s_i + a_i > 0$ and 1 otherwise. A higher attack intensity increases the probability of a breach, $\frac{d\delta_i}{da_i} \geq 0$, while higher security lowers it, $\frac{d\delta_i}{ds_i} \leq 0$.

[9] In an extended version of the model, if an attacker can receive non-pecuniary benefits (e.g., from the disruption of transactions), one could interpret r as a sum of financial gains and non-pecuniary benefits (i.e., a high value of r represents a combination of a high degree of financial and non-financial motivations). This interpretation may be particularly relevant for some state-sponsored cyber attacks who benefit from the disruption in other states' services.

We assume that firm security investment is observable only to firms and attackers, owing to their relative sophistication, but not to clients. In reality, clients often have limited access to reliable information about firms' cyber security practices. For example, publicly traded companies may report total spending on IT infrastructure or cyber security in their financial reports, but rarely provide granular data.

Clients. A mass M of identical clients indexed by m have exogenous transaction needs with value-at-risk V_m at $t = 2$. Our approach effectively assumes that the demand for transactions is insensitive to cyber risk and focuses our attention to its supply-side impact.

Clients simultaneously allocate their transactions across firms based on the transaction value at risk, where $v_{im} \geq 0$ is the value-at-risk of client m with firm i and $V_m = \sum_{i=1}^{N} v_{im}$. We normalize $V_m \equiv 1$ without loss of generality. Thus, the total market size for all firms is M and the transaction value-at-risk at firm i is $V_i \equiv \int_0^M v_{im}\,\mathrm{d}m$. Each client maximizes her expected utility

$$U_m = \sum_{i=1}^{N}(1 - f_i - \delta_i \ell_i)v_{im}, \tag{4}$$

where the client enjoys the transaction value-at-risk net of fees and the loss upon a successful breach. We view the first term of Equation 4 as the cyber risk-adjusted net return $(1-f_i-\delta_i\ell_i)$ per transaction value-at-risk v_{im} allocated to each firm i.

3 Benchmarks

We derive the equilibrium with unobservable and observable security investment as well as the constrained-efficient allocation. All of these allocations serve as benchmarks for the subsequent equilibrium with signalling as well as the evaluation of regulatory measures.

3.1 Unobservable security investment

We start by defining the equilibrium and then characterize it. A key aspect is that clients do not observe the security investment of firms and form beliefs about it, $\mu(s_i) = \widehat{s}_i$.

Definition 1 (Equilibrium with unobservable security investment.) *An equilibrium is given by ℓ_i^*, a_i^*, s_i^*, f_i^*, $\widehat{s}_i$, and v_{im}^* for all $i = 1, ..., N$ and $m \in [0, M]$, where:*

1. *At $t = 3$, the attack strategy on firm i is $a(V_i, s_i, \ell_i) = \arg\max_{a_i} \pi_A$ for any V_i, s_i, ℓ_i.*
2. *At $t = 2$, client beliefs about security investment of firm i are $\widehat{s}_i = \arg\max_{s_i} \pi_i(\mathbf{f})$ for any fees $\boldsymbol{f} \equiv \{f_i\}_{i=1}^N$ and attack modality $\boldsymbol{l} \equiv \{\ell_i\}_{i=1}^N$, given $a(V_i, s_i, \ell_i)$.*
3. *At $t = 2$, the transaction allocation strategy is $v_{im}(\widehat{s}_i, \mathbf{f}) = \arg\max_{v_{im}} U_m$ s.t. $\sum_{i=1}^N v_{im} = V_m$ for any fees $\boldsymbol{f}$ and attack modality $\boldsymbol{l}$, given $\widehat{s}_i$ and $a(V_i, s_i, \ell_i)$.*
4. *At $t = 1$, $(\mathbf{s}^*, \mathbf{f}^*)$ is a Nash equilibrium among firms. That is, $(s_i^*(\ell_i), f_i^*(\ell_i)) = \arg\max_{s_i, f_i} \pi_i$ for any ℓ_i, given the choices of the other firms (s_{-i}, f_{-i}), the allocation strategies and beliefs of clients, $v_{im}(\widehat{s}_i, \mathbf{f})$ and $\widehat{s}_i$, and the attack strategies $a(V_i, s_i, \ell_i)$.*
5. *At $t = 0$, the attack modality is $\boldsymbol{l}^* = \arg\max_{\ell_i} \pi_A$, given $(\mathbf{s}^*, \mathbf{f}^*)$, allocation strategies and beliefs, $v_{im}(\widehat{s}_i, \mathbf{f})$ and $\widehat{s}_i$, and the attacker's own future attack strategies, $a(V_i, s_i, \ell_i)$.*

We focus on symmetric equilibria, which requires that (i) all firms invest identically in security, $s_i^* = s^*$, and offer identical fees, $f_i^* = f^*$; (ii) clients allocate $\int_0^M v_{im}^* = \frac{M}{N}$ to each firm, and (iii) the attacker chooses the same intensity and modality on all firms, $a_i^* = a^*$, and $\ell_i^* = \ell^*$. This equilibrium is summarized in Proposition 1 and proven in Appendix B.1.

Proposition 1 (Unobservable security investment.) *In equilibrium, the attacker chooses an attack modality that only impacts the firm's clients, $\ell^* = 1$. Firms do not invest in security, $s^* = 0$, charge no fees, $f^* = 0$, and cyber attacks succeed with certainty, $\delta^* = 1$.*

We provide some intuition for the equilibrium when firms' security investment is unobservable (and cannot be credibly communicated to clients). Attackers choose to attack whenever the security investment is low enough. Clients form beliefs about firm security investments and choose firms that offer the highest return net of the fee and adjusted for cyber risk. Firms face Bertrand-style competition and set fees to break even in expectation, as they would otherwise not attract any business. Finally, attackers choose to target clients (and not firms), e.g. via ransomware attacks. In equilibrium, firms have no incentives of investment in cyber security. This breakdown in the market for cyber security provision by firms results from the attacker's modality choice. If the attacker were to focus on the firm, it would protect its own assets by investing in security. By targeting clients' assets and data instead, the firm lacks the incentive to invest in security. This is not because clients would not prefer it—and pay for the service through higher fees—but because firms cannot credibly convince clients that they have in fact invested on their behalf.

3.2 Observable Security Investment

How would equilibrium outcomes change when security investment is publicly observable? We next evaluate the benchmark (BM) of symmetric information (whereby clients also observe firm security investment). Appendix B.2 contains the definition of equilibrium, which is similar to Definition 1 but without client beliefs about security investment, and all derivations. We have the following characterization of the equilibrium.

Proposition 2 (Observable security investment.) *Symmetric equilibria exist. They are characterized by* $v_m^{BM} = \frac{M}{N}$ *and*

$$s^{BM} = \begin{cases} \frac{rM}{N} & \text{if } 2rc \leq 1 \\ \frac{M}{4c^2 rN} & \text{if } 2rc > 1, \end{cases} \tag{5}$$

and

$$a^{BM} = \begin{cases} 0 & \text{if } 2rc \leq 1 \\ \frac{M}{2cN}\left(1 - \frac{1}{2rc}\right) & \text{if } 2rc > 1. \end{cases} \tag{6}$$

The attack modality is indeterminate, $\ell_i^{BM} = [0, 1]$*. For a given* ℓ_i^{BM}*, the equilibrium fee is*

$$f_i^{BM}(\ell_i^{BM}) = \begin{cases} rc & \text{if } 2rc \leq 1 \\ 1 - \ell_i^{BM} + \frac{2\ell_i^{BM} - 1}{4rc} & \text{if } 2rc > 1. \end{cases} \tag{7}$$

There are two cases, with the first case ($2rc \leq 1$) referring to an equilibrium with no successful attacks and the second case ($2rc > 1$) referring to one with successful attacks happening with positive probability. This is clearest when considering the equilibrium probability of a successful attack:

$$\delta^{BM} = \begin{cases} 0 & \text{if } 2rc \leq 1 \\ 1 - \frac{1}{2rc} & \text{if } 2rc > 1. \end{cases} \tag{8}$$

To build some intuition, note that firms can credibly convey observable security investments to clients. Consequently, firms are willing to protect clients when attackers focus attacks on client assets, because firm can pass these costs on to clients via higher fees (without fearing the loss of business). Hence, both security investment s^{BM} and fees f^{BM} are positive in equilibrium for all ℓ^{BM}. Intuitively, the choice of attack modality can no longer induce firms to invest little in security, as it did under unobservable security investment. Finally, the higher security investment under symmetric information reduces firms' attack vulnerability compared to unobservable security investment (see Equation 8). To summarize:

Corollary 1 (Attack vulnerability and investment transparency.) *There are fewer expected breaches when cyber security investment is observable,* $\delta^{BM} < \delta^*$*.*

3.3 Constrained-efficient allocation

The breakdown in the market for security investment with asymmetric information has a strong negative impact on welfare, as we show next. To make welfare statements, we consider a social planner who maximizes utilitarian welfare W that aggregates firm profits and client utility (so the fee payments wash out):

$$W \equiv \sum_{i=1}^{N} \left((1 - \delta_i) V_i - c s_i \right), \tag{9}$$

where we assume that the planner does not take into account the utility of the attacker.[10] In this second-best problem, the planner cannot choose the actions of the attacker. That is, a (constrained) planner takes the attack strategy of the attacker as given (as do private agents). We use the second-best allocation as our welfare benchmark throughout the paper.

One can show that the second-best outcome (SB) is identical to the symmetric information benchmark:

$$s^{SB} = s^{BM}, \qquad a^{SB} = a^{BM}. \tag{10}$$

When, however, security investment is unobservable to clients, firms do not invest in security, $s^* = 0$, and attacks succeed with certainty, $\delta^* = 1$, reducing welfare to zero. $W^* = 0$. Thus, the unobservability of firm security investment reduces welfare dramatically.

Proposition 3 (Welfare in the Benchmarks.) *Under the symmetric information benchmark, the equilibrium outcomes s^{BM} and a^{BM} achieve the social planner's second-best welfare outcome. When clients do not observe security investment, $s^* = 0$, and welfare is zero.*

Proposition 3 raises the natural question as to how one might affect equilibrium outcomes to improve welfare in a world where security investment is unobservable. To this end, we

[10] One can view this assumption in two ways. As a proxy for i) the attacker being a foreign agent, as many cyber attackers are, where the planner intends only to maximize the welfare of its domestic constituents, thereby placing no value on the welfare of agents outside of the home country. Or ii) the social value that a society places on the welfare of those who gain through criminal means.

evaluate a menu of private and public options to encourage the level of or the transparency about security investment in the remainder of the paper.

4 Costly signalling with independent security ratings

We begin by considering a market-based solution, whereby firms can purchase a credible signal to inform clients of their security investment. We denote all variables in this section with the subscript 'R' to indicate security ratings are available, thereby distinguishing them from the unobservable investment benchmark which has no subscript, e.g., $s_{R,i}$ denotes firm i security investment when security rating signals are available. We assume that firms may (individually) purchase a signal when they invest in security and choose fees at $t = 1$, and that this signal comes at a cost of $\kappa s_{R,i}$ which preserves the linearity of the model (and thus tractability). As a tie-breaker, we assume that if a firm has the option of a signalling or non-signalling strategy that would earn identical client shares, it prefers the latter.

This costly signal may reflect the security ratings offered by firms such as BitSight and UpGuard. They offer independent cybersecurity evaluation services and package the results into easily-digestible ratings, similar to credit scores.[11] The prevalence of independent security rating firms shows the costly nature of cybersecurity disclosure: firms may increase vulnerability if they permit clients to view their cybersecurity operations directly, and they may face concerns of "window-dressing" for any information disclosed themselves. In fact, Amir et al. (2018) show that firms under-report information on cyber attacks due to managerial incentives to withhold negative information, as investors cannot usually uncover evidence of these attacks on their own. The proliferation of industry "Cyber Security Working Groups" also highlights the firms' desire to publicly showcase their investment in cyber security.[12]

[11] A testament to their value, the Canadian Government partnered with cyber security ratings firm SecurityScorecard in 2024 to assign letter-grades to critical infrastructure firms as an assessment of their cyber resilience (Bloomberg 2023).

[12] Some examples of organizations with cyber security working groups include CITA (Wireless), EFCOG (Energy), and ABA (Banking).

We model these scores as follows. At $t = 2$, clients observe the signal of security investment for each firm i, $\theta_i \in \{\varnothing, s_{R,i}\}$, where $\theta_i = \varnothing$ indicates no signal from firm i, and $\theta_i = s_{R,i}$ signals firm i's investment perfectly (for simplicity). Clients then choose either firms who signal or firms who do not, depending on which offers them a higher utility. A firm that signals can incentivize clients to choose it for their transactions if it can set $f_{R,i}$ and $s_{R,i}$ such that client utility exceeds that of non-signalling firms. These (signalling) firms choose $f_{R,i}$ and $s_{R,i}$ to maximize their profit:

$$\pi_{R,i} = f_{R,i}V_{R,i} - (1 - \ell_{R,i})\delta_{R,i}V_{R,i} - cs_{R,i} - \kappa s_{R,i}, \tag{11}$$

subject to attracting a positive market share. Unlike in Section 3.1, the best other offer may come from signalling or non-signalling firms. Solving the problem by working backwards, we arrive at the following result, which is proven in Appendix B.4.

Proposition 4 (Costly signalling.) *The attacker targets both firms and clients, $\ell_R^* \in (0, 1)$, in a way that no firm uses the signalling technology in equilibrium. We have:*

$$\ell_R^* = \begin{cases} \sqrt{1 - 4rc[1 - r(c + \kappa)]} & \text{if } 2r(c + \kappa) \leq 1, \\ \sqrt{\frac{\kappa}{c+\kappa}} & \text{if } 2r(c + \kappa) > 1. \end{cases} \tag{12}$$

Firms choose a security investment s_R^* such that $0 < s_R^* < s^{SB}$. Welfare is improved over the case where investment is unobservable, but does not attain the second-best benchmark.

Figure 1: Attack Modality Choice with Costly Signaling

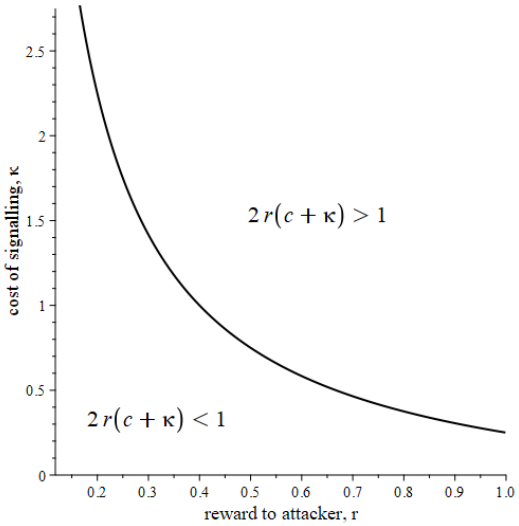

Figure 1 shows the two regions of attack modality choice. In the upper region ($2r(c+\kappa) > 1$), attackers can earn positive profits regardless of whether firms signal. In the lower region, attackers can only earn a positive profit if firms do not signal. In both cases, the attacker chooses an attack modality ℓ_R^* such that the firm chooses not to signal. Parameter: $c = 0.5$.

Figure 1 shows the attack modality choice ℓ_R^* and its two equilibrium regions. For $2r(c+\kappa) \leq 1$, attackers only earn a positive profit when firms do not signal their security investment. In this region, attacker profit is zero when firms signal, but the attacker's choice of ℓ_R^* encourages firms not to. The attacker does so by offering a value of ℓ_R^* low enough such that the firms' incentive to invest in security outweighs the benefits from signalling. The value $\ell_R^* = \sqrt{1 - 4rc[1 - r(c+\kappa)]}$ is the highest value of ℓ such that clients would be better off if the the firms were not to signal. For $2r(c+\kappa) > 1$, the attacker is guaranteed positive profit regardless of whether the firm signals. Nonetheless, the attacker is better off when the firm chooses not to signal. As in the first region, the attacker chooses the highest value of ℓ such that clients are better off when firms do not signal, which yields $\ell_R^* = \sqrt{\frac{\kappa}{c+\kappa}}$.

Proposition 4 highlights an important tension in the attacker's modality choice when signalling is available. First, if the attacker focuses their attacks on firms directly, this induces firms to invest in security to protect themselves from losses regardless of whether

they signal. If, however, the attacker focuses entirely on clients—as they do when security investment is unobservable and signalling is unavailable—then clients would earn higher utility from a firm that chooses to signal. Here, the effect of competition between firms kicks in, inducing firms to signal to capitalize on clients' preference for security signalling.

The equilibrium outcome rests in between the observable and unobservable security benchmarks. The attacker chooses a higher ℓ_R^* than in the unobservable security benchmark case to prevent firms from signalling, improving firm security investment as they guard themselves against a greater percentage of direct attacks towards firms. Thus, when signaling is available, firms invest more in security and attackers succeed less often, resulting in higher welfare. The second-best welfare level, however, cannot be reached using costly signalling alone, as a percentage of attacks still target clients and security investment remains unobservable in equilibrium. The security investment chosen by a non-signalling firm would only be equal to the second-best level if the attacker chose to exclusively attack firms directly ($\ell_R^* = 0$), but doing so would be akin to the observable security benchmark in which $\ell^* = 0$, as firms have full liability for losses in both cases.[13] Instead, the attacker faces a 'balancing act' between targeting firms to prevent security investment that follows when firms are induced to signal, and targeting clients to reduce direct security investment by firms.

In equilibrium with security ratings, firms choose a level of security investment $s_{R,i}^* = s_R^*$:

$$s_R^* = \frac{(1-\ell_R^*)^2 M}{4c^2 r N} \tag{13}$$

where ℓ_R^* is as in Equation 12. While security investment always improves over the unobservable benchmark (Section 3.1), it falls below the observable security benchmark (Section 3.2), as the attacker's modality choice ensures that firms: a) do not invest enough to fully secure when $2rc \leq 1$, and b) reduce security investment by a factor of $(1-\ell_R^*)^2$ when $2rc > 1$.

[13] Whether it is even feasible for signalling firms to achieve the second-best allocation depends on whether the cost from signalling creates an additional welfare losses or is simply a transfer. In any case, our result shows that even if signalling entails only private costs and no social costs, second-best cannot be attained because of the incentives of the attacker in choosing its attack modality.

This results in the following vulnerability to cyber attacks:

$$\delta_R^* = 1 - \frac{(1-\ell_R^*)}{2rc} \tag{14}$$

We summarize the firm vulnerability discussion in the following proposition.

Proposition 5 (Costly signalling and attack vulnerability.) *The availability of costly signals reduces the attack vulnerability but does not reach the second-best benchmark:*

$$\delta^{BM} < \delta_R^* < \delta^*. \tag{15}$$

4.1 Testable implications

To derive testable implications of the model, we consider the comparative statics of the equilibrium with costly signalling. We examine three sets of model parameters: the reward captured by the attacker r, the marginal cost of security c, and the marginal cost of a signalling κ. We summarize these comparative statics in the following proposition, with the corresponding derivations found in Appendix B.5.

Proposition 6 (Comparative statics under signalling.) *The effects of changes in parameters* $\left(\frac{M}{N}, r, c, \kappa\right)$ *on equilibrium outcomes* $(\delta_R^*, a_R^*, s_R^*, f_R^*, \ell_R^*)$ *are given in Table 1, where arrows indicate increasing or decreasing in the specified parameter.*

	M/N	r	c	κ
Vulnerability (δ_R^*)	0	n.m	n.m	↑
Attack intensity (a_R^*)	↑	n.m	n.m	↓
Security Investment (s_R^*)	↑	n.m	n.m	↓
Fees (f_R^*)	0	↑	↑	↓
Attack Modality (l_R^*)	0	↓	↓	↑

Table 1: Comparative statics of firm vulnerability (δ_R^*), attack intensity (a_R^*), security investment (s_R^*), fees (f_R^*), and attack modality (ℓ_R^*) for parameters market tightness (M/N), reward to the attacker (r), cost of security investment (c), and cost of signalling (κ) in the case with costly signalling. The notation "n.m." refers to "non-monotonic".

Proposition 6 highlights three key testable implications: the impact of cost parameters on attacker modality and firm fees, and the impact of signalling costs on breaches.

Testable Implication 1 (Attack Modality.) *For firms or industries with a higher* (r, c) *or lower* κ*, attackers focus a higher percentage of their attacks on firms directly (lower* ℓ_R^**).*

Signalling reduces attacker profits, so attackers shift their attack toward firms to reduce the value of signalling. Thus, in industries where the cost of signalling (κ) is higher, the degree to which attackers shift their attacks to firms is reduced, as the cost of signalling itself reduces signalling their security investment to clients.

For industries in which transactions are more valuable to the attacker (higher r), attackers have an incentive to increase their attacking intensity, all else equal. Hence, clients have a greater preference for firms to signal their investment. To dissuade firms from signalling, attackers respond by focusing a higher percentage of their attacks on firms. Similarly, if firms face a higher cost of security (c), they reduce their investment in security. Client preference for signalling increases, leading attackers to shift their attacks to firms as a counterbalance.

Testable Implication 2 (Breaches.) *Firms or industries with higher cost of signalling* κ *experience more breaches on average,* δ_R^**.*

When signals for security investment are available, industries or firms for which signalling is marginally more expensive should expect to be more vulnerable to attacks. The core driver of this implication follows from Testable Implication 1, as the increased cost of signalling deters firms from signalling their investment and, ceteris paribus, encouraging attackers to target clients directly to a greater degree. Lacking the same incentives to provide enhanced security when attackers shift their focus to clients, security investment is also lower in these industries. The result is even more beneficial to attackers, as they achieve greater chances of success with lower attack intensity.

Testable Implication 3 (Client Fees.) *Firms or industries with a higher* (r, c) *or lower* κ *charge higher client fees* f_R^*.

Testable Implication 3 highlights the correlation between client fees and i) attacker incentives to attack (r) and, ii) disincentives for firms to provide security (c). When either of these parameters are high, firms face higher costs to providing security either directly (c) or through direct attacker incentives to increase intensity (r). Perhaps counter-intuitively, our model also implies that firms or industries with higher costs of signalling charge *lower* fees. These predictions arise from the impacts of (r, c, κ) on attack modality: when attackers focus a greater percentage of their attacks on the firm directly, fees increase.[14]

5 Regulatory measures

A market-based response to the issue of unobservable security investment, via signalling, improves welfare but does not achieve the second-best level of welfare. Hence, there may be scope for regulatory interventions to improve efficiency. How regulatory intervention incentivizes firms is crucial: Curti et al. (2023) examine the efficacy of "data breach notification laws" that require firms to publicly notify following successful data breaches. They find that such laws fail to improve cyber security by firms, and argue that laws that enforce a minimum industry standard of cyber security would be more effective. Our approach follows this line of thought, focusing on policies that require action on cyber security either by direct mandate, or via a 'skin-in-the-game' incentive. In particular, we consider two regulatory policies in this section: i) minimum levels of security investment, and; ii) consumer protection measures.[15]

[14] The notable exception to this correlation is that fees always increase in r, whereas attack modality (ℓ^*) is independent of r when $2r(c + \kappa) > 1$.

[15] Because our model focuses on firms attacked in isolation, these regulatory interventions are also aimed at individual firms, which provides a lower bound on the benefits from regulation. In practice, there may be substantial network externalities, increasing the benefits from regulation by preventing successful attacks from spilling over to other firms.

5.1 Minimum investment in security

Consider a regulator who imposes a minimum security investment standard on firms at the beginning of $t = 0$ (that is, before the attacker chooses an attack modality). Hence, firm security investment decisions now face the constraint $s_i \geq s_{MIN}$. Moreover, the investment floor s_{MIN} is publicly observable, informing clients that firms cannot invest below the floor.

Under symmetric information, a floor for security investment is either irrelevant or inefficient, as Section 3.3 shows: the equilibrium security investment reaches the second-best welfare level.[16] Under asymmetric information, however, welfare is below the second-best level, suggesting that a floor on security investment may improve welfare. In fact, our model implies that an appropriately calibrated minimum investment level achieves the second-best outcome, as summarized in the following proposition (proven in Appendix B.6.)

Proposition 7 (Minimum security investments.) *The socially optimal level of minimum security investment is $s^*_{MIN} = s^{SB}$. Then, firms choose $s^{MS} = s^{SB}$ and choose not to signal their security investment. The attacker's modality choice is irrelevant and the second-best level of welfare is achieved (despite asymmetric information).*

This allocation not only maximizes welfare but also client utility, subject to zero profits of firms when they pay no signalling costs. Thus, a firm that simply chooses $s^* = s_{MIN}$ and does not pay signalling costs offers the utility-maximizing contract to clients.

5.2 Consumer protection

When clients cannot observe firm security investment, the division of losses between firms and clients (i.e., attack modality) following a successful attack plays an important role in both the attacker's profit-maximization problem and the firm's security investment decision.

[16] A minimum investment requirement is irrelevant if it is below the investment chosen in the unregulated equilibrium. When the requirement exceeds this level, there is inefficient (over-)investment in security, above what customers are willing pay for given the value derived from transacting and their potential losses.

In particular, attackers have the incentive to ensure clients bear some or all of the losses so that firms under-invest in security relative to the symmetric information benchmark, even when the option to signal investment is available, as previously shown.

In practice, do consumers bear losses associated with successful cyber attacks? The 2017 data breach of credit reporting agency Equifax led to a fine of $575m for a data breach affecting 147 million consumers, of which $425m was dedicated to the settlement with consumers, which is roughly $2.89 per affected consumer.[17] Comparatively, in the same year, the average cost of identity theft was $1038 (CNBC 2017). A consumer protection regulator may thus be concerned with consumer losses from cyber attacks, and how such a regulation would affect welfare.

We next investigate whether a regulator that redistributes losses back to firms upon a successful cyber attack provides the incentive for firms to increase security investment beyond the signalling equilibrium level, and whether the result is welfare-improving. To capture these issues, we consider a regulator who can impose a penalty p on firms that are breached. This penalty takes the form of a transfer from firms to their clients, based on the losses suffered by clients. Specifically, when client m suffers a loss of $\ell_i v_{im}$ upon a breach, the firm is fined a total of $p\ell_i v_{im}$, which is transferred directly to the client. We also assume that the regulator acts again at the beginning $t = 0$, before the attacker chooses their modality ℓ.

The penalty changes the payoffs of agents in the model. The non-signalling firm's profit function changes to

$$\pi_i = f_i V_i - [1 - (1 - p)\ell_i]\delta_i V_i - cs_i, \tag{16}$$

for any given penalty p. Similarly, a signalling firm's profit function changes to

$$\pi_i = f_i V_i - [1 - (1 - p)\ell_i]\delta_i V_i - (c + \kappa)s_i. \tag{17}$$

[17] See Federal Trade Commission (2022) for settlement details and ArsTechnica (2019) for an example of media coverage.

And client utility changes to

$$U_m = \sum_{i=1}^{N} [1 - f_i - \delta_i(1-p)\ell_i] v_{im}, \tag{18}$$

We have the following result on consumer protection (see Appendix B.7 for a proof).

Proposition 8 (Consumer protection.) *If a regulator makes firms liable for all client losses, $p = 1$, investment in security becomes independent of the attack modality and the regulator can achieve the second-best level of welfare (even under asymmetric information).*

Consumer protection regulation that obliges firms to reimburse clients for losses associated with cyber attacks improves not only client utility but also achieves the second-best level of welfare. When a regulator makes firms fully liable for client losses, $p^* = 1$, the non-signalling firm sets its security at the second-best level, $s^{CP} = s^{SB}$, regardless of the attack modality. A signalling firm is unable to achieve this outcome, as the costly signal lowers its optimal security investment. Client utility is strictly higher at non-signalling firms, and no clients chooses the signalling firms. Thus, in the case where firms are fully liable for cyber attacks, they invest in security equal to the second-best established in Section 3.3.

Finally, we acknowledge that while a regulatory solution to impose liability on firms may be necessary in some industries, in practice, some firms may voluntarily offer their clients financial protection from attacks. For example, in some jurisdictions, credit card issuers will reimburse consumers if their cards are used for fraudulent purposes.[18]

6 Conclusion

As digital transactions increase in importance, firms play a critical role in safeguarding their clients' assets and data. Thefts of cryptocurrencies from exchanges (BitMart), data breaches

[18]The results for the case in which deep-pocketed firms can choose a publicly observable liability parameter, rather than having one imposed by a regulator, are qualitatively similar to the regulator's solution.

of credit rating agencies (Equifax), and the extortion of firms that provide essential banking services (ICBC) highlight the breadth and severity of cyber attacks. However, the complex and opaque nature of cyber security investment reduces firm incentives to mitigate the risks.

In this paper, we propose a model of cyber risk and cyber security investment. Clients and firms are subject to cyber attacks and firms invest in security to protect their and their clients' assets and data. The key friction is the unobservability of firm security investment by clients due to its opaque and private nature, giving rise to a principal-agent problem. This leads firms to under-invest in security, thereby facing greater vulnerability to cyber attacks and negatively impacting welfare when compared to a second-best benchmark in which information about security investment is easily obtained and understood by clients.

A market-based solution is for firms to purchase a credible signal of their security investment, as offered by cyber security ratings firms such as BitSight and UpGuard. Our model predicts that the availability of these services reduces firm vulnerability to cyber attacks and improves welfare compared to an economy not served by cyber security ratings firms.

Since access to cyber security ratings does not reach the second-best outcome, we examine whether regulatory measures could improve welfare further. First, targeting investment directly via minimum security standards can accomplish this task, reaching the second-best outcome. Alternatively, a regulator may protect clients from losses by imposing liability on firms upon successful attacks. By redistributing the cost of attacks from clients to firms, this type of "skin-in-the-game" consumer protection policy incentivizes firms to invest in security for themselves. We conclude that client protection regulation effectively resolves the principal-agent problem and the regulated economy reaches the second-best outcome.

References

Acemoglu, Daron, Azarakhsh Malekian, and Asu Ozdaglar, 2016, Network security and contagion, *Journal of Economic Theory* 166, 536–585.

Akey, Pat, Stefan Lewellen, Inessa Liskovich, and Christoph Schiller, 2021, Hacking corporate reputations, Technical report.

Amir, Eli, Shai Levi, and Livne Tsafrir, 2018, Do firms underreport information on cyber attacks? evidence from capital markets., *Review of Accounting Studies* 23, 1177–1206.

Anand, Kartik, Chanelle Duley, and Prasanna Gai, 2022, Cybersecurity and financial stability, *Deutsche Bundesbank Discussion Paper No 08/2022* .

Anderson, Ross, 2001, Why information security is hard-an economic perspective, in *Proceedings 17th Annual Computer Security Applications Conference (ACSAC), 2001*, 358–365, IEEE.

Anderson, Ross, Chris Barton, Rainer Böhme, Richard Clayton, Michel JG Van Eeten, Michael Levi, Tyler Moore, and Stefan Savage, 2013, Measuring the cost of cybercrime, in *The economics of information security and privacy*, 265–300.

Anderson, Ross, and Tyler Moore, 2006, The economics of information security, *Science* 314, 610–613.

ArsTechnica, 2019, You're probably not going to get your $125 from the Equifax settlement, https://arstechnica.com/tech-policy/2019/07/youre-probably-not-going-to-get-your-125-from-the-equifax-settlement/, Accessed: 2023-10-19.

August, T., and T.I Tunca, 2006, Network software security and user incentives, *Management Science* 52, 1703–1720.

Becker, Gary S., 1968, Crime and punishment: An economic approach, *Journal of Political Economy* 76, 169–217.

Berkman, henk, Jonathan Jona, Gladys Lee, and Naomi Soderstrom, 2018, Cybersecurity awareness and market valuations., *Journal of Accounting and Public Policy* 37, 508–526.

Bloomberg, 2021, Hackers said to seize $150 million from BitMart exchange, https://www.bloomberg.com/news/articles/2021-12-05/hackers-said-to-seize-150-million-from-bitmart-crypto-exchange, Accessed: 2023-10-19.

Bloomberg, 2023, Canada will use letter grades to assess companies' cyberresilience, https://www.bloomberg.com/news/articles/2024-01-11/canada-will-use-letter-grades-to-limit-danger-from-hackers?srnd=premium-canada, Accessed: 2024-01-15.

Carnegie, 2022, Carnegie endowment for international peace: Timeline of cyber incidents involving financial institutions, https://carnegieendowment.org/specialprojects/protectingfinancialstability/timeline, Accessed: 2023-10-19.

CNBC, 2017, Identity theft, fraud cost consumers more than $16 billion, https://www.cnbc.com/2017/02/01/consumers-lost-more-than-16b-to-fraud-and-identity-theft-last-year.html, Accessed: 2023-10-19.

CNN, 2024, JPMorgan Chase fights off 45 billion hacking attempts each day, https://edition.cnn.com/2024/01/17/investing/jpmorgan-fights-off-45-billion-hacking-attempts-each-day/index.html, Accessed: 2024-01-18.

Crosignani, Matteo, Marco Macchiavelli, and André F. Silva, 2023, Pirates without borders: The propagation of cyberattacks through firms' supply chains, *Journal of Financial Economics* 147, 432–448.

Curti, Filippo, Ivan Ivanov, Marco Macchiavelli, and Tom Zimmermann, 2023, City hall has been hacked! the financial costs of lax cybersecurity .

Duffie, Darrell, and Joshua Younger, 2019, *Cyber runs* (Brookings).

Dynes, Scott, Eric Goetz, and Michael Freeman, 2007, Cyber security: Are economic incentives adequate?, in *International Conference on Critical Infrastructure Protection*, 15–27.

Dziubiński, Marcin, and Sanjeev Goyal, 2013, Network design and defence, *Games and Economic Behavior* 79, 30–43.

ECB, 2024, ECB to stress test banks' ability to recover from cyberattack, https://www.bankingsupervision.europa.eu/press/pr/date/2024/html/ssm.pr240103~a26e1930b0.en.html, Accessed: 2024-06-24.

Eisenbach, Thomas M, Anna Kovner, and Michael Junho Lee, 2022, Cyber risk and the US financial system: A pre-mortem analysis, *Journal of Financial Economics* 145, 802–826.

Federal Trade Commission, 2022, Equifax data breach settlement, https://www.ftc.gov/enforcement/refunds/equifax-data-breach-settlement, Accessed: 2023-10-19.

Florackis, Chris, Christodoulos Louca, Roni Michaely, and Michael Weber, 2023, Cybersecurity risk, *Review of Financial Studies* 36, 351–407.

FT, 2020, BA fine for data breach in 2018 sharply reduced, https://www.ft.com/content/910188fc-863e-45f2-8dfd-8fe7e9fe9ca0, Accessed: 2023-10-19.

Garg, Priya, 2020, Cybersecurity breaches and cash holdings: Spillover effect, *Financial Management* 49, 503–519.

Gordon, Lawrence A, and Martin P Loeb, 2002, The economics of information security investment, *ACM Transactions on Information and System Security (TISSEC)* 5, 438–457.

Goyal, Sanjeev, and Adrien Vigier, 2014, Attack, Defence, and Contagion in Networks, *The Review of Economic Studies* 81, 1518–42.

Hoyer, Britta, and Kris de Jaegher, 2016, Strategic network disruption and defense, *Journal of Public Economic Theory* 18, 802–830.

Jamilov, Rustam, Hélène Rey, and Ahmed Tahoun, 2022, The anatomy of cyber risk, *NBER Working Paper No. w28906* .

Kamiya, Shinichi, Jun-Koo Kang, Jungmin Kim, Andreas Milidonis, and René M. Stulz, 2021, Risk management, firm reputation, and the impact of successful cyberattacks on target firms, *Journal of Financial Economics* 139, 719–749.

Kopp, Emanuel, Lincon Kaffenberger, and Christopher Wilson, 2017, Cyber risk, market failures, and financial stability, Technical report.

Kotidis, Antonis, and Stacey L. Schreft, 2022, Cyberattacks and financial stability: Evidence from a natural experiment, Technical report.

Reuters, 2023, Gang says icbc paid ransom over hack that disrupted us treasury market, https://www.reuters.com/technology/cybersecurity/icbc-paid-ransom-after-hack-that-disrupted-markets-cybercriminals-say-2023-11-13/, Accessed: 2024-01-15.

Securities and Exchange Commission, 2022, Statement on proposal for mandatory cybersecurity disclosures, https://www.sec.gov/news/statement/gensler-cybersecurity-20220309, Accessed: 2023-10-19.

Vásquez, Jorge, 2022, A theory of crime and vigilance, *American Economic Journal: Microeconomics* 14, 255–303.

A Notation

Exogenous Parameters	
Parameters	Definition
V	Total client transaction value-at-risk
r	Reward of the attacker
M	Mass of clients
N	Number of firms
c	Marginal cost of security investment by firms
κ	Marginal cost of signalling per unit of security investment
p	Penalty imposed by regulator for successful attacks (under consumer protection policy)
Endogenous Variables	
Variable	Definition
δ_i	Probability of a successful cyber attack at firm i (so firm i is breached)
v_{im}	Transaction value-at-risk of client m of transactions allocated to firm i
V_i	Total transaction value-at-risk of transactions allocated to firm i
a_i	Attack intensity by the cyber attacker on firm i
f_i	Fee per transaction value-at-risk charged by firm i
ℓ_i	Attack modality chosen by cyber attacker against firm i
s_i	Security investment of firm i
$\widehat{s}_i$	Belief about security investment of firm i formed by client, $\mu(s_i) = \widehat{s}_i$
ζ	Best offer made by firm $j \neq i$
π_i	Profit of firm i
U	Utility of client
π_A	Profit of cyber attacker
θ_i	Signal of cyber security rating that takes a value of $\{\varnothing, s_i\}$
Variable Identifiers	
Identifier	Definition
$*$	Equilibrium with unobservable security investment, e.g s^*
BM	Benchmark allocation with observable investment, e.g. s^{BM}
SB	Second-best allocation, e.g. s^{SB}
R	Equilibrium with unobservable security investment but cyber security ratings are available
$\mu(s)$	Belief over security investment

B Proofs

B.1 Unobservable security investment (Proposition 1)

We solve for an equilibrium by working backwards.

Attack intensity. At $t = 3$ the attacker chooses a_i to maximize π_A that can be written as:

$$\pi_A = r \sum_{i=1}^{N} \left(\frac{a_i}{a_i + s_i} V_i - a_i \right), \tag{19}$$

whenever $\delta_i > 0$. The first derivative with respect to a_i is set equal to zero, so $\frac{s_i}{(a_i+s_i)^2} r V_i - 1 = 0$. Imposing $a_i \geq 0$ yields the bound on security investment, $\overline{s}_i$. The outcome $\delta_i = 0$ occurs for all $s_i > \overline{s}_i$. Alternatively, solving for a_i yields the branch for $\delta_i > 0$. Taken together, we have:

$$a(V_i, s_i, \ell_i) \equiv \begin{cases} \sqrt{s_i r V_i} - s_i & \text{if } s_i \leq r V_i \equiv \overline{s}_i \\ 0 & \text{if } s_i > \overline{s}_i. \end{cases} \tag{20}$$

Intuitively, the attacker chooses to attack firm i with positive intensity when their yield of firm i's transaction value-at-risk, rV_i, is high relative to the security level s_i.

Client beliefs and allocation of transactions across firms. With security investment unobservable by clients at $t = 2$, clients form beliefs about security investment at each firm, $\widehat{s}_i$, conditional upon which they allocate their transactions across all firms. Consistent with sequential rationality, we assume clients believe that each firm chooses security to maximize their profit, given observable fees f_i and attack modality ℓ_i. That is, clients assign beliefs $\widehat{s}_i = \arg\max_{s_i} \pi_i(\mathbf{f})$ at $t = 2$. (Perfect Bayesian Equilibrium does not impose a structure on off-equilibrium beliefs, so many beliefs about off-equilibrium f_i and ℓ_i are possible. We assume that clients hold these beliefs regardless of whether they observe an on-equilibrium or off-equilibrium f_i and ℓ_i.) These beliefs enter client utility in Equation 4 through the probability of a successful attack against firm i, $\delta_i(\widehat{s}_i)$.

Conditional on beliefs, each client maximizes her utility by allocating her transactions among the firms offering the highest value, which is determined by the highest cyber risk-adjusted return net of fees (Equation 4), that is $[1 - f_i - \ell_i \delta(V_i, \widehat{s}_i)]$. The client allocates zero to all other firms.

Security investment and fees. At $t = 1$, each firm i chooses s_i and f_i to maximize the expected profits in Equation 2, taking as given the allocation strategy of clients, $v_{im}(\mathbf{s}, \mathbf{f}, \ell)$, and the attacker's intensity strategy, $a(V_i, s_i, \ell_i)$, for all attack type choices ℓ. As a result of the allocation strategy, each firm is subject to a positive-market-share constraint:

$$(1 - f_i - \delta(V_i, s_i, \ell_i)\ell_i) \geq \zeta, \tag{21}$$

where $\zeta = \max_{j \neq i} (1 - \delta(V_j, s_j, \ell_j)\ell_j - f_j)$ is the best offer made by other firms. Competition for positive market share leads to a Bertrand-style "race-to-the-bottom" competition in fees: firms undercut each other to capture the market until each firm earns zero expected profits in equilibrium.

At $t = 1$, each firm i is aware of what client beliefs $\widehat{s}_i$ are at $t = 2$. Therefore, a firm that chooses s_i to maximize its profit for a given f_i chooses $\widehat{s}_i = s^*$ in equilibrium, which is given by:

$$\widehat{s}_i = s^* = \begin{cases} \frac{rM}{N} & \text{if } 2rc \leq 1 - \ell \\ \frac{(1-\ell)^2 M}{4c^2 rN} & \text{if } 2rc > 1 - \ell. \end{cases} \tag{22}$$

Firms have no incentive to choose a value of $s^* \neq \widehat{s}_i$ because client's believe that they are choosing their profit-maximizing value given observables, $\widehat{s}_i = \arg\max_{s_i} \pi_i(\mathbf{f})$. Next, we obtain $\delta(V_i, s^*, \ell)$,

$$\delta(\ell) = \begin{cases} 0 & \text{if } 2rc \leq 1 - \ell \\ 1 - \frac{(1-\ell)}{2rc} & \text{if } 2rc > 1 - \ell. \end{cases} \tag{23}$$

The equilibrium value f^* is then determined by the zero profit assumption after substituting in s^* and $\delta(\ell)$:

$$f^*(\ell) = \begin{cases} rc & \text{if } 2rc \leq 1 - \ell \\ 1 - \ell - \frac{(1-\ell)^2}{4rc} & \text{if } 2rc > 1 - \ell. \end{cases} \tag{24}$$

Attack modality. Since security investment decreases in ℓ, the attacker's expected payoff decreases in s, giving an incentive to target the attacks on clients (higher ℓ), encouraging the firm to reduce s. Hence, the optimal attack modality is $\ell^* = 1$. At $\ell^* = 1$, firms then choose $s^* = 0$ and the attacker succeeds with certainty while selecting $a^* = 0$ and incurring no costs.

B.2 Observable security investment (Proposition 2)

Here we define and characterize the equilibrium of our model under symmetric information.

Definition 2 (Observable security investment.) *An equilibrium is given by ℓ_i^*, a_i^*, s_i^*, f_i^*, and v_{im}^* for all $i = 1, ..., N$ and $m \in [0, M]$ and is found via backward induction:*

1. *At $t = 3$, the attack strategy on firm i is $a(V_i, s_i, \ell_i) = \arg\max_{a_i} \pi_A$, for any V_i, s_i, ℓ_i.*
2. *At $t = 2$, the transaction allocation strategy is $v_{im}(\mathbf{s}, \mathbf{f}, \boldsymbol{l}) = \arg\max_{v_{im}} U_m$ subject to $\sum_{i=1}^N v_{im} = V_m$ and the attack strategies $a(V_i, s_i, \ell_i)$, for any $(\mathbf{s}, \boldsymbol{f}, \boldsymbol{l}) \equiv \{f_i, s_i, \ell_i\}_{i=1}^N$.*[19]
3. *At $t = 1$, $(\mathbf{s}^*, \mathbf{f}^*)$ is a Nash equilibrium among firms. That is, $(s_i^*(\ell_i), f_i^*(\ell_i)) = \arg\max_{s_i, f_i} \pi_i$, for any ℓ_i, given the choices of the other firms (s_{-i}, f_{-i}), the allocation strategies of clients $v_{im}(\mathbf{s}, \mathbf{f}, \boldsymbol{l})$, and the attack strategies $a(V_i, s_i, \ell_i)$.*

[19] Note that $V_i(\mathbf{s}, \mathbf{f}, \mathbf{l}) = \int_0^M v_{im'}(\mathbf{s}, \mathbf{f}, \mathbf{l})\, dm'$ is independent of m because each client has zero mass.

4. *At $t = 0$, the attack modality $\boldsymbol{\ell}^* = \arg\max_{\ell_i} \pi_A$, given the Nash equilibrium among the firms $(\mathbf{s}^*, \mathbf{f}^*)$, the allocation strategies of clients $v_{im}(\mathbf{s}, \mathbf{f}, \boldsymbol{\ell})$, and the attacker's own future attack strategies $a(V_i, s_i, \ell_i)$.*

We continue to focus on symmetric equilibria, which requires that (i) all firms invest identically in security, $s_i^* = s^*$, and offer identical fees, $f_i^* = f^*$; (ii) clients allocate $\int_0^M v_{im}^* = \frac{M}{N}$ to each firm, and (iii) the attacker chooses the same attack intensity and modality, $a_i^* = a^*$, and $\ell_i^* = \ell^*$.

Attack intensity. The attacker's problem does not change by allowing clients to observe firm investment. Thus, their best response function is still the solution $a(V_i, s_i, \ell_i)$ in Equation 20.

Allocation of transactions across firms. At $t = 2$ the clients allocate their transactions based on transaction value-at-risk v_{im} such that $v_{im}(\mathbf{s}, \mathbf{f}) = \arg\max_{v_{im}} U_m$ with $\sum_{i=1}^{N} v_{im} = V_m$ for any s_i and f_i and given $a(V_i, s_i)$. Clients do not need to form beliefs about s_i, as they just observe it. Client utility is:

$$U_m = \sum_{i=1}^{N} [1 - f_i - \ell_i \delta(V_i, s_i)] v_{im}. \tag{25}$$

A client's utility is maximized by allocating $v_{im} > 0$ to any group of firms with the highest $[1 - f_i - \ell_i \delta(V_i, s_i)]$. Clients equally allocate v_{im} amongst these firms.

Security investment and fees. At $t = 1$, each firm i chooses s_i and f_i to maximize the expected profits in Equation 2, taking as given the allocation strategy of clients, $v_{im}(\mathbf{s}, \mathbf{f}, \ell)$, and the attacker's intensity strategy, $a(V_i, s_i, \ell_i)$, for all attack modality choices ℓ subject to attracting a positive market share characterized by Equation 21. Competition for positive market share leads to a Bertrand-style "race-to-the-bottom" in fees where firms undercut each other to capture the market until each firm earns zero expected profits in equilibrium. Hence, given the optimal client transaction allocation $v_m^* = \frac{M}{N}$, we solve the firm's constrained optimization problem.

At $t = 1$, each firm i assumes that amongst the other firms $-i$, a different firm j offers the highest value of $[1 - f_i - \ell_i \delta(V_i, s_i)]$, where $[1 - f_j - \ell_j \delta(V_j, s_j)] = \zeta$ and $\zeta > 0$. Each firm chooses s_i and f_i to maximize its profits, taking the actions of the other firms (s_{-i}, f_{-i}) as given, such that $[1 - f_i - \ell_i \delta(V_i, s_i)] \geq \zeta$. The simplified first-order conditions with respect to s_i and f_i are

$$\frac{\partial \pi_i}{\partial s_i} : \frac{(1 - \ell_i)}{2} \sqrt{\frac{V_i}{s_i r}} - c + \lambda \frac{\ell_i}{2} \sqrt{\frac{1}{V_i s_i r}} = 0, \tag{26}$$

$$\frac{\partial \pi_i}{\partial f_i} : V_i - \lambda = 0, \tag{27}$$

where λ is a Lagrange multiplier. It can be shown that $\lambda = 0$ implies $\zeta < 0$, which violates the clients' participation constraints. Thus, $\lambda = V_i$. Solving for s_i^* and inputting V_i, we obtain:

$$s^* = \begin{cases} \frac{rM}{N} & \text{if } 2rc \leq 1 \\ \frac{M}{4c^2 rN} & \text{if } 2rc > 1, \end{cases} \tag{28}$$

Next, we impose symmetry among all firms and invoke a zero profit condition to solve for ζ^* and f^*. Substituting s^* into $a^*(\cdot)$ yields

$$a^* = \begin{cases} 0 & \text{if } 2rc \leq 1 \\ \frac{M}{2cN}\left(1 - \frac{1}{2rc}\right) & \text{if } 2rc > 1. \end{cases} \tag{29}$$

which we use in conjunction with s^*, to obtain $\delta(a^*, s^*)$ from Equation 3:

$$\delta(s^*, a^*) = \begin{cases} 0 & \text{if } 2rc \leq 1 \\ 1 - \frac{1}{2rc} & \text{if } 2rc > 1. \end{cases} \tag{30}$$

Finally, to solve for $f^*(\ell)$, we solve $\pi(a^*, s^*, f) = 0 = fV - (1-\ell)\delta(a^*, s^*)V - cs^*$.

$$f^*(\ell) = \begin{cases} rc & \text{if } 2rc \leq 1 \\ 1 - \ell + \frac{2\ell - 1}{4rc} & \text{if } 2rc > 1. \end{cases} \tag{31}$$

The equilibrium is piece-wise, depending on whether $a^* > 0$, and are shown in Equations 28–31.

Attack modality. Finally, at $t = 0$ the attacker chooses an attack modality ℓ_i for each firm, taking as given the firms' security and fee decisions $s_i(\ell_i)$ and $f_i(\ell_i)$, the allocation strategy of clients, $v_{im}(\mathbf{s}, \mathbf{f}, \ell)$, and the attacker's own intensity strategy, $a(V_i, s_i, \ell_i)$. Since the firm security investment strategies s^* is independent of ℓ_i^*, changes in the allocation of losses between the firms and their clients does not incite a response from either firms or clients, leaving the attacker's profit function unaffected. Hence, any attack modality, $\ell_i^* \in [0, 1]$ is optimal.

Proposition 9 (Comparative statics for observable security investment.) *The effects of changes in parameters* $\left(\frac{M}{N}, r, c\right)$ *on equilibrium outcomes* $(\delta^*, a^*, s^*, f^*)$ *are given in Table 2, where arrows indicate increasing or decreasing in the specified parameter.*

	M/N		r		c		
	$2rc \leq 1$	$2rc > 1$	$2rc \leq 1$	$2rc > 1$	$2rc \leq 1$	$rc \in (0.5,1)$	$rc \geq 1$
Vulnerability (δ^*)	0		0	↑	0	↑	
Attack intensity (a^*)	0	↑	0	↑	0	↑	↓
Security Investment (s^*)	↑		↑	↓	0	↓	
Fees (f^*)	0		↑	↓	↑	↓	

Table 2: Comparative statics of firm vulnerability (δ^*), attack intensity (a^*), security investment (s^*) and fees (f^*) for parameters market tightness (M/N), value of the asset to the attacker (r), and cost of security investment (c) for observable security investment.

Proof. Recall the functions for $(s^*, a^*, \delta^*, f^*)$ from Equations 28–31, respectively. Most of the comparative statics follow by inspection. For s^*, for example, we see that, for $2rc \leq 1$, s^* is

independent of c and increases in M/N and r. For $2rc > 1$, s^* increases in M/N but decreases in c and r. Consider f^* next. For all parameters, f^* is independent of M/N. If $2rc \leq 1$, f^* increases in c and r, but decreases in both parameters when $2rc > 1$.

Next, for a^*, for $2rc \leq 1$, we can see that a^* is independent of all parameters. For $2rc > 1$, a^* increases in M/N and in r. To study how a^* changes in c, we take the first derivative:

$$\frac{\partial a^*}{\partial c} = -\frac{M(rc-1)}{2c^3 rN} \tag{32}$$

Hence, $\frac{\partial a^*}{\partial c}$ increases in c for $rc \in (1/2, 1)$, and decreases in it for $rc > 1$. Finally, for $2rc \leq 1$, δ^* is independent of all parameters. For $2rc > 1$, δ^* is independent of M/N but increases in r and c.

B.3 Second-best allocation (Proposition 3)

First, the social planner cannot choose a^*, but instead can choose s^{SB}, f^{SB}, and V^{SB}. Because f^{SB} does not enter the welfare function and $a^{SB}(s^{SB})$ is a function of s, we optimize over the welfare function by choosing s_i. The optimal s^{SB} is thus determined by the first order condition:

$$\frac{\partial W}{\partial s} = \frac{\sqrt{V}}{2\sqrt{sr}} - c = 0 \tag{33}$$

$$\iff s^{SB} = \begin{cases} rV_i & \text{if } 2rc \leq 1 \\ \frac{V_i}{4rc^2} & \text{if } 2rc > 1. \end{cases} \tag{34}$$

Where the case of $2rc < 1$ admits from s^* solving $a(s) = 0$. Because this is the same optimal s_i for all i, the social planner chooses $V^{SB} = V_i = M/N$ for all i (though any choice produces equal welfare). Equation 33 implies $s^{SB} = s^{BM}$. Then, because the attacker chooses a^{SB} based on s^{SB}, it must be that $a^{SB}(s^{SB}) = a^{SB}(s^{BM}) = a^{BM}$. Hence, the second-best welfare outcome is identical to market outcome under symmetric information.

Under asymmetric information (clients cannot observe security investment), the equilibrium outcome is $s^* = 0$, which implies $\delta_i = 1$ for all i. Hence:

$$W(s_i = 0) = \sum_{i=1}^{N} \left((1 - \delta_i(s_i = 0)) V_i - c \times 0 \right) = \left((1-1) V_i - 0 \right) = 0 \tag{35}$$

B.4 Costly signalling (Propositions 4 and 5)

In this case, we denote security investment for firm i as $s_{R,i}$. At $t = 2$, clients continue to assign beliefs $\mu(s_{R,i}) = \widehat{s}_{R,i}$ to the security investment of firm i that does not signal (i.e., $\theta_i = \varnothing$). When firms signal, the signal is perfect and thus $\mu(s_R) = s_{R,i}$. Clients then divide their transactions

equally among all firms that offer the highest utility of either signalling $[1 - f_{R,i} - \ell_{R,i}\delta(V_{R,i}, s_{R,i})]$ or non-signalling ($\theta_i = \varnothing$), $[1 - f_{R,i} - \ell_{R,i}\delta(V_{R,i}, \widehat{s}_{R,i})]$.

The firm's problem at $t = 1$ has two parts. First, the firm solves its problem twice: assuming that it does not signal, and assuming that signals. Second, the firm decides whether to signal or not, based on which option will offer higher utility to clients and thus earn the firm a share of their business. Firm i signals if and only if the clients' utility is higher under a signalling contract than a non-signalling contract $[1 - f_{R,i} - \ell_{R,i}\delta(V_{R,i}, s_{R,i})] > [1 - _{R,i} - \ell_{R,i}\delta(V_{R,i}, \widehat{s}_{R,i})]$.

A firm that signals chooses $s_{R,i}(\theta_i = s_{R,i})$ and $f_{R,i}(\theta_i = s_{R,i})$ to maximize its profit in Equation 11. Focusing on symmetric equilibria, all firms arrive at identical solutions, and thus clients divide their transactions among all firms equally, such that each firm's share is $V = \frac{M}{N}$. This allows us to drop the firm subscript i in what follows. The solutions are given by:

$$s_R(\theta = s_R) = \begin{cases} \frac{rM}{N} & \text{if } 2r(c+\kappa) \leq 1 \\ \frac{M}{4Nr(c+\kappa)^2} & \text{if } 2r(c+\kappa) > 1. \end{cases} \tag{36}$$

$$f_R(\ell_R, \theta = s_R) = \begin{cases} r(c+\kappa) & \text{if } 2r(c+\kappa) \leq 1 \\ 1 - \ell_R + \frac{2\ell_R - 1}{4r(c+\kappa)} & \text{if } 2r(c+\kappa) > 1. \end{cases} \tag{37}$$

Then, following from $s_R(\theta = \varnothing)$ and $a_R(\theta = \varnothing)$, we have $\delta_R(\theta_i = s_R)$ for completeness:

$$\delta_R(\theta = s_R) = \begin{cases} 0 & \text{if } 2r(c+\kappa) \leq 1 \\ 1 - \frac{1}{2r(c+\kappa)} & \text{if } 2r(c+\kappa) > 1. \end{cases} \tag{38}$$

For firms that do not signal, clients continue to form beliefs $\mu(s_R) = \widehat{s}_R$. The non-signalling firm's problem and solutions are then identical to those in Proposition 1.

Since firms continue to earn zero profits, the entire cost of signalling is passed on to their clients. Thus, when firms select to signal, client utility is

$$U(\theta = s_R) = \begin{cases} \frac{M}{N(1-r(c+\kappa))} & \text{if } 2r(c+\kappa) \leq 1 \\ \frac{M}{4Nr(c+\kappa)} & \text{if } 2r(c+\kappa) > 1. \end{cases} \tag{39}$$

If firms do not signal, the security and fee choices (for a given ℓ_R^*) result in a client utility of

$$U(\ell_R, \theta = \varnothing) = \begin{cases} \frac{M(1-rc)}{N} & \text{if } 2rc \leq 1 - \ell_R \\ \frac{\frac{M}{N}(1-\ell_R)(1+\ell_R)}{4rc} & \text{if } 2rc > 1 - \ell_R. \end{cases} \tag{40}$$

If the firm signals, the attacker is unable to influence its security investment through the attack modality ℓ_R. Thus, at $t = 0$, the attacker considers its problem in two parts, depending on the

hypothetical response of a firm that signals. First, if $2r(c+\kappa) \leq 1$, a firm which chooses to signal will choose $s_R(\theta = s_R) = \frac{rM}{N}$. Were the firm to signal, the attacker's best response would be $a_R(\theta = s_R) = 0$, resulting in $\pi_A(\theta = s_R) = 0$. If the firm does not signal, the firm's security response $s_R(\theta = \varnothing)$ decreases in ℓ_R. The attacker then chooses the highest value of ℓ_R such that firms would choose not to signal, which is to say $[1 - f_R^* - \ell_R \delta(s_R^*) \mid \theta = s_R^*] > [1 - f_R^* - \ell_R \delta(\widehat{s}) \mid \theta = \varnothing]$, where $f_R^*(\theta = \varnothing)$ is given by Equation 24. The solution ℓ_R^* is given by the first segment of Equation 12. The attacker earns a profit of $\pi_A > 0$, so it always prefers to induce no signalling by the firm. Second, if $2r(c+\kappa) > 1$, a firm which chooses to signal will choose $s_R(\theta = s_R) < \frac{rM}{N}$. Were the firm to signal, the attacker's optimal response at $t = 3$ would result in a realized profit of:

$$\pi_A(\theta = s_R) = \frac{M}{N}\left(\sqrt{r} - \frac{1}{2(c+\kappa)\sqrt{r}}\right)^2 \tag{41}$$

As before, the attacker can ensure the firm does not signal if it chooses an ℓ_R that satisfies $[1 - f_R^* - \ell_R \delta(s_{R,i}^*)] > [1 - f_R^* - \ell_R \delta(\widehat{s})]$. In this case, the highest value of ℓ_R that satisfies this condition is given by the second segment of Equation 12. The attacker's profit, for any ℓ_R were the firm not to signal is given:

$$\pi_A(\theta = \varnothing) = \frac{M}{N}\left(\sqrt{r} - \frac{(1-\ell_R)}{2c\sqrt{r}}\right)^2 \tag{42}$$

Inserting $\ell_R = \sqrt{\frac{\kappa}{c+\kappa}}$ from Equation 12, the payoff in Equation 42 is greater than in Equation 41 for any $c > 0$ and $\kappa > 0$. Thus, the attacker prefers to induce no signalling by the firm.

Given the equilibrium modality choice by the attacker ℓ_R^* to induce firms not to, this has implications for their optimal security investment and fee decision. Particularly, the full security investment threshold that is in terms of $\ell_R^* = 1 - 2rc$. We show that for either ℓ_R^* about the threshold $2r(c+\kappa) = 1$, it must be that $\ell_R^* > 1 - 2rc$, and thus the firm never invests such that the fully secure under the optimal modality choice by the attacker.

First, let $2r(c+\kappa) \leq 1 \Rightarrow \ell_R^* = \sqrt{1 - 4rc(1-rc) + 4r^2c\kappa}$. To show that $\ell_R^* > 1 - 2rc$, suppose not. Moreover, suppose $\kappa = 0$ to impose a smallest possible ℓ_R^*. Then rearranging leads to:

$$\sqrt{1 - 4rc(1-rc)} \leq 1 - 2rc \iff \sqrt{(1-2rc)^2} \leq 1 - 2rc \tag{43}$$

Then, for any $\kappa > 0$, it must be that the left-hand side is large, a contradiction. Hence, $2r(c+\kappa) > 1 \Rightarrow \ell_R^* > 1 - 2rc$. Next, let $2r(c+\kappa) > 1 \Rightarrow \ell_R^* = \sqrt{\frac{\kappa}{c+\kappa}}$. To show that $\ell_R^* > 1 - 2rc$, suppose not. Moreover, suppose $\kappa = \frac{1}{2r} - c$, which is the smallest possible κ such that $2r(c+\kappa) > 1$ to impose the smallest possible ℓ_R^*. Then rearranging leads to:

$$\sqrt{\frac{\left(\frac{1}{2r} - c\right)}{c + \left(\frac{1}{2r} - c\right)}} \leq 1 - 2rc \iff \sqrt{1 - 2rc} \leq 1 - 2rc \tag{44}$$

Then, because ℓ_R^* has a lower bound of 0, if $2rc > 1$, then for any ℓ_R^*, $2rc > 1 - \ell_R^*$. A contradiction. Thus, $2r(c+\kappa) > 1 \Rightarrow \ell_R^* > 1 - 2rc$. With this property, we can thus write (s_R^*, f_R^*):

$$s_R^* = \begin{cases} \frac{(1-\sqrt{1-4rc(1-rc)+4r^2c\kappa})^2 M}{4c^2 rN} & \text{if } 2r(c+\kappa) \leq 1 \\ \frac{(1-\sqrt{\frac{\kappa}{c+\kappa}})^2 M}{4c^2 rN} & \text{if } 2r(c+\kappa) > 1. \end{cases} \tag{45}$$

$$f_R^* = \begin{cases} 1 - \sqrt{1-4rc(1-rc)+4r^2c\kappa} - \frac{(1-\sqrt{1-4rc(1-rc)+4r^2c\kappa})^2}{4rc} & \text{if } 2r(c+\kappa) \leq 1 \\ 1 - \sqrt{\frac{\kappa}{c+\kappa}} - \frac{(1-\sqrt{\frac{\kappa}{c+\kappa}})^2}{4rc} & \text{if } 2r(c+\kappa) > 1. \end{cases} \tag{46}$$

And firm vulnerability δ_R is given by:

$$\delta_R^* = \begin{cases} 1 - \frac{(1-\sqrt{1-4rc(1-rc)+4r^2c\kappa})}{2rc} & \text{if } 2r(c+\kappa) \leq 1 \\ 1 - \frac{(1-\sqrt{\frac{\kappa}{c+\kappa}})}{2rc} & \text{if } 2r(c+\kappa) > 1. \end{cases} \tag{47}$$

We turn to the proof of Proposition 5. When security is unobservable, $\delta^* = 1$, from Proposition 1. In the benchmark, we have that vulnerability is piecewise either $\delta^{BM} = 0 < \delta^* = 1$. If $2rc > 1$ instead, then $\delta^{BM} = 1 - \frac{1}{2rc}$ (Equation 30), so $\delta^{BM} < \delta^* = 1$. Hence, $\delta^{BM} < \delta^* = 1$ for all (r, c, M, N). With costly signalling, no firms signal, and δ_R^* is as in Equation 47, which can be rearranged as:

$$\delta_R^* = 1 - \frac{1}{2rc} + \frac{\ell_R^*}{2rc} = \delta^* + \frac{\ell_R^*}{2rc} \tag{48}$$

Then, since $\ell_R^* \in (0, 1)$, it must be that $\delta^{BM} \leq \delta_R^* < \delta^*$ for any $(r > 0, c > 0, \kappa > 0)$.

B.5 Comparative statics with costly signalling (Proposition 6)

In the signalling equilibrium, attackers choose ℓ_R such that firms will not signal. Thus, the equilibrium values of (s_R^*, f_R^*) are as in Equations 45 and 46. For the value of δ_R^*, the equation is as in Equation 47. a_R^* takes Equation 20 and inserts s_R^* from Equation 45 to arrive at:

$$a_R^* = \sqrt{\frac{(1-\ell_R)^2 M^2}{4c^2 rN^2}} - \frac{(1-\ell_R^*)^2 M}{4c^2 rN} \tag{49}$$

where ℓ_R^* is characterized piecewise by Equation 12, given by:

$$\ell_R^* = \begin{cases} \sqrt{1-4rc[1-r(c+\kappa)]} & \text{if } 2r(c+\kappa) \leq 1, \\ \sqrt{\frac{\kappa}{c+\kappa}} & \text{if } 2r(c+\kappa) > 1. \end{cases} \tag{50}$$

Thus, the dynamics of our key equilibrium values of Proposition 6 may differ on either side of the threshold $2r(c+\kappa) = 1$.

First, we sign the first-order condition (FOC) of ℓ_R^* in $\left(\frac{M}{N}, r, c, \kappa\right)$, as we will need this in later steps. By inspection, the FOC in $\frac{M}{N}$ is zero.

$$\begin{aligned}
\frac{\partial \ell_R^*}{\partial r} &= \begin{cases} \frac{2c(2r(c+\kappa)-1)}{\sqrt{1-4rc[1-r(c+\kappa)]}} < 0 & \text{if } 2r(c+\kappa) \leq 1, \\ 0 & \text{if } 2r(c+\kappa) > 1. \end{cases} \\
\frac{\partial \ell_R^*}{\partial c} &= \begin{cases} \frac{2r(r(2c+\kappa)-1)}{\sqrt{1-4rc[1-r(c+\kappa)]}} < 0 & \text{if } 2r(c+\kappa) \leq 1, \\ -\frac{\kappa}{2\sqrt{\frac{\kappa}{c+\kappa}}(c+\kappa)^2} < 0 & \text{if } 2r(c+\kappa) > 1. \end{cases} \\
\frac{\partial \ell_R^*}{\partial \kappa} &= \begin{cases} \frac{2cr^2}{\sqrt{1-4rc[1-r(c+\kappa)]}} > 0 & \text{if } 2r(c+\kappa) \leq 1, \\ \frac{c}{2\sqrt{\frac{\kappa}{c+\kappa}}(c+k)^2} > 0 & \text{if } 2r(c+\kappa) > 1. \end{cases}
\end{aligned} \tag{51}$$

Much of the signing of Equation 51 follows by inspection. For the numerator of $\frac{\partial \ell_R^*}{\partial r}$ when $2r(c+\kappa) \leq 1$, this condition implies that the numerator must be negative. Similarly, if $2r(c+\kappa) \leq 1$, then it must also be true that $(r(2c+\kappa)-1) < 0$, as $r(2c+\kappa) < 2r(c+\kappa)$. Thus, $\frac{\partial \ell_R^*}{\partial c} < 0$.

Next, it follows by inspection that (δ_R^*, f_R^*) are independent of M/N, and (a_R^*, s_R^*) are increasing in M/N. Taking FOC of s_R^* and f_R^* in (r, c, κ) respectively, we find that:

$$\begin{aligned}
\frac{\partial s_R^*}{\partial r} &= -\frac{(1-\ell_R^*)M}{4rc^2N}\left(2\frac{\partial \ell_R^*}{\partial r} + \frac{(1-\ell_R^*)}{r}\right) : n.m. \\
\frac{\partial s_R^*}{\partial c} &= -2\frac{(1-\ell_R^*)M}{4rc^2N}\left(\frac{\partial \ell_R^*}{\partial c} + \frac{(1-\ell_R^*)}{c}\right) : n.m. \\
\frac{\partial s_R^*}{\partial \kappa} &= -\frac{(1-\ell_R^*)M}{2rc^2N}\frac{\partial \ell_R^*}{\partial \kappa} < 0
\end{aligned} \tag{52}$$

$$\begin{aligned}
\frac{\partial f_R^*}{\partial r} &= \frac{\partial \ell_R^*}{\partial r}\left(\frac{(1-\ell_R^*)}{2rc} - 1\right) + \frac{(1-\ell^*)^2}{4r^2c} > 0 \\
\frac{\partial f_R^*}{\partial c} &= \frac{\partial \ell_R^*}{\partial c}\left(\frac{(1-\ell_R^*)}{2rc} - 1\right) + \frac{(1-\ell_R^*)^2}{4rc^2} > 0 \\
\frac{\partial f_R^*}{\partial \kappa} &= \left(\frac{(1-\ell_R^*)}{2rc} - 1\right)\frac{\partial \ell_R^*}{\partial \kappa} < 0
\end{aligned} \tag{53}$$

Recall that $2rc > 1 - \ell_R^*$ and $\frac{\partial \ell_R^*}{\partial \kappa} > 0$. Moreover, because $\frac{\partial \ell_R^*}{\partial r} \geq 0$, we can sign the above FOCs as they appear. For $\frac{\partial s_R^*}{\partial r}$, where $\frac{\partial \ell_R^*}{\partial r}$ is zero, the value is negative. If, however, $\frac{\partial \ell_R^*}{\partial r}$ is negative, then we show graphically that the equation is non-monotonic. Similarly, when $2r(c+\kappa) > 1$, the second term of $\frac{\partial \ell_R^*}{\partial c}$ simplifies to $\frac{\partial \ell_R^*}{\partial c} = \frac{1}{c} - \frac{\frac{3\kappa c}{2} - \kappa}{(c+\kappa)^2\sqrt{\frac{\kappa}{\kappa+c}}c}$, which is positive by graphical inspection.

Moreover:

$$\begin{aligned}
\frac{\partial a_R^*}{\partial r} &= -\frac{1}{2}\left(\frac{(1-\ell_R^*)}{4c^2}\right)^{-\frac{1}{2}}\left(\frac{(1-\ell_R^*)}{2c^2}\frac{\partial \ell_R^*}{\partial r}\right) - \frac{\partial s_R^*}{\partial r} : n.m. \\
\frac{\partial a_R^*}{\partial c} &= \left(\frac{1}{2}\left(\frac{s_R^* r M}{N}\right)^{-\frac{1}{2}} r\frac{M}{N} - 1\right)\frac{\partial s_R^*}{\partial c} : n.m. \\
\frac{\partial a_R^*}{\partial \kappa} &= \left(\frac{1}{2}\left(\frac{s_R^* r M}{N}\right)^{-\frac{1}{2}} r\frac{M}{N} - 1\right)\frac{\partial s_R^*}{\partial \kappa} < 0
\end{aligned} \tag{54}$$

Given the FOCs below, inspection yields that the dynamics of δ_R^* in (r, c, κ) are inversely proportional to the dynamics in s_R^*.

$$\begin{aligned}
\frac{\partial \delta_R^*}{\partial r} &= -\frac{1}{2}\left(\frac{s_R^* N}{rM}\right)^{-\frac{1}{2}}\frac{N}{rM}\frac{\partial s_R^*}{\partial r} \propto \frac{1}{s_R^*} \\
\frac{\partial \delta_R^*}{\partial c} &= -\frac{1}{2}\left(\frac{s_R^* N}{rM}\right)^{-\frac{1}{2}}\frac{N}{rM}\frac{\partial s_R^*}{\partial c} \propto \frac{1}{s_R^*} \\
\frac{\partial \delta_R^*}{\partial \kappa} &= -\frac{1}{2}\left(\frac{s_R^* N}{rM}\right)^{-\frac{1}{2}}\frac{N}{rM}\frac{\partial s_R^*}{\partial \kappa} > 0
\end{aligned} \tag{55}$$

B.6 Minimum security standards (Proposition 7)

The second-best allocation s^{SB} can be shown to be the value of s_i that maximizes client utility subject to a non-signalling firm's zero-profit condition. Thus, were security investment observable to clients, no non-signalling firm could offer a security-fee schedule that would improve client utility.

Clients beliefs are modified to $\widehat{s}_i = \arg\max_{s_i} \pi_i(\mathbf{f})$ s.t. $s_i \geq s_{MIN}$, so as to reflect the new constraint. Suppose a non-signalling firm offers f^* given by Equation 7. The solution to the firms' problem, subject to the constraint $s_i \geq s_{MIN}$ is $s^{MS} = s_{MIN} = s^{SB}$, where 'MS' indicates the minimum security level. A signalling firm cannot offer a better contract to clients, as it must charge higher fees for any given s_i. However, without observable security investment, such a firm could conceivably deviate to another level of security that satisfies the constraint, while continuing to offer the same fees. Were this possible, $s^* = s_{MIN}$ would not be an equilibrium.

We consider this possible deviation for the two ranges of parameters in turn. First, when $2rc \leq 1$ and thus $s_{MIN} = \frac{rM}{N}$, any security value above s_{MIN} results only in higher costs to the firm with no security benefits since $\delta^* = 0$, and thus positive deviation is neither profitable to the firm, nor desirable to clients. A non-signalling firm which offers f^* given by Equation 7 and investing $s^* = s_{MIN}$ would capture all client volume from any non-signalling firm pursuing an alternative strategy.

Second, the case when $2rc > 1$ and thus $s_{MIN} = \frac{M}{4rc^2N}$, is somewhat more complex, as the firm could increase its security investment. This would increase its costs but would also lower the probability of a successful attack. Consider that the fee which would earn a firm zero-profit if it selected $s^* = s_{MIN}$ is given by Equation 7. When clients do not observe security investments, their beliefs are $\widehat{s}_i = \arg\max_{s_i} \pi_i(\mathbf{f})$ s.t. $s_i \geq s_{MIN}$. This value is given by $s_i = \frac{(1-\ell^*)^2 V_i}{4rc^2}$ which is less than s_{MIN} for any $\ell^* > 0$. Firm profitability is decreasing for all values above $s_i = \frac{(1-\ell^*)^2 V_i}{4rc^2}$, and the firm chooses the minimum value which satisfies the constraint, s_{MIN}. Thus, as in the first case, a non-signalling firm which offers f^* given by Equation 7 and investing $s^* = s_{MIN}$ would capture all client volume from any non-signalling firm pursuing an alternative strategy.

When the regulator sets $s_{MIN} = s^{SB}$, signalling firms cannot offer an advantage over non-signalling firms. Were they to choose the same level of security as a non-signalling firm, they would incur higher costs and thus pass on higher fees to clients, with no added security investment. Further, the level of security investment that maximizes client utility for a signalling firm is both below the minimum security level and results in utility of less than the second-best. Any deviations above the minimum for a signalling firm only further erodes both client utility. Thus, a signalling firm can offer no improvement to clients and would capture no client volume.

B.7 Consumer protection (Proposition 8)

Assigning the value $p = 1$ transforms a non-signalling firm's profit function given by Equation 16, such that with respect to s_i, it is identical to the social planner's welfare function given by Equation 9. Thus, this firm chooses $s^{CP} = s^{SB}$, regardless of whether clients view its security investment, while the signalling firm's security investment remains equal to Equation 36. As noted in the proof of Proposition 7, this is also the outcome that maximizes client utility subject to the firm's zero-profit condition. A signalling-firm with a profit function given by Equation 17 is unable to recreate this outcome for any $\kappa > 0$ and thus receives no transactions $V_i = 0$.