# A Growth and Innovation Model of the Modern Data Economy

Orlando Gomes[*], Roxana Mihet[†] and Kumar Rishabh[‡§]

July 26, 2023

**Abstract**

In this paper, we formulate a growth model for the data economy, considering the dual role of data as a vital business optimization tool and a potential target for cybercrime, with the associated risks of theft and destruction. We explore the dynamic interplay between cybercrime risk, digital innovation, and their repercussions on economic growth. Unequivocally, cybercrime results in lower stocks of knowledge, lost productivity and lower growth for all firms in the economy. The silver-lining is that cybercrime risk encourages firms to pursue digital innovation that boosts productivity in other domains. We observe a 3% escalation in R&D activities, a 5% rise in patenting activity, and a 0.4% increase in the diversity of patents filed in response to a one-standard deviation shock in cyber risk. Additionally, we demonstrate that our findings predominantly apply to data-intensive firms, whereas non-data-intensive firms do not exhibit increased general innovation in response to cyber security threats.

**Keywords**: Data economy, data theft, data breaches, cyber-risk, growth, artificial intelligence, innovation.

**JEL-Codes:** D8, O3, O4, G3, L1, L2, M1

[*]Lisbon Accounting and Business School, ISCAL. Contact: omgomes@iscal.ipl.pt

[†]Swiss Finance Institute at HEC Lausanne, and CEPR. Contact: roxana.mihet@unil.ch

[‡]University of Lausanne, and University of Basel. Contact: kumar.rishabh@unil.ch

[§]First version: January 31, 2023. This version: July 26, 2023. We are indebted to Chris Florackis, Christodoulos Louca, Roni Michaely, and Michael Weber for sharing their data on cyberrisk with us. We thank participants at various conferences, including the EEA, SFI Annual Days 2023, the Economics of ICT 2023, and ERMAS 2023 for useful feedback. We are also grateful for insightful suggestions from Laura Veldkamp and Tarun Ramadorai. Roxana Mihet acknowledges generous funding for this project from The Sandoz Family Foundation - Monique de Meuron Programme. All authors declare no conflicts of interest related to this project. All errors are our own. Send all correspondence to: roxana.mihet@unil.ch.

# 1 Motivation

The cost of business data breaches and theft can be significant and long-lasting. For firms, cybercrime can lead to financial losses, loss of sensitive information, lost productivity, reputational damage, legal consequences, and decreased customer trust. Cyberattacks can also disrupt business operations and result in costly downtime. Additionally, the resources spent on preventing and mitigating cybercrime could be put towards other investments that could drive economic growth. For societies, the cost of cybercrime goes beyond financial losses, and can range from compromised infrastructure, to national security issues, to missed economic opportunities. In the last 10 years in the United States, the monetary damage caused by reported cybercrime increased 12 fold from $ 581 million in 2012 to $ 6.9 billion in 2022 (Internet Crime Complaint Center).[1] As cybercrime is becoming costlier, more frequent and more aggressive over time, regulators are worried that it could harm U.S. companies' ability to remain leaders in innovation globally.

In this paper, we develop a framework to study the interactions between cybercrime and digital innovation, together with the individual and combined effects of these phenomena on economic growth. We first build a growth model of the data economy in which data is information that helps firms optimize their business processes and is subject to cyberrisk, meaning that it can be damaged and destroyed by cyber criminals. We use the framework to quantify the impact of cyberrisk on firm growth and innovation. Cybercrime results in lower stocks of knowledge, lost productivity and lower growth for all firms in the economy. With innovation, long-run sustained growth remains achievable even with cybercrime, because lower stocks of knowledge due to cyberrisk can be compensated by an increasing number of available product varieties. Firms can hedge against increased cyberrisk by innovating more to create alternative sources of data, which makes data even more valuable for cyber criminals and results in a cybercrime driven innovation loop.

We empirically quantify this feedback loop and find that firms experience a 3% increase in R&D, a 5% increase in patenting activity, and 0.4% expansion in the diversity of the patent fields in response to a one-standard deviation shock in cyberrisk, control-

[1] While accurately estimating the total cost of cybercrime is difficult because the costs comprise not only of criminal revenue and direct losses, but also indirect losses and defense costs (Anderson et al. (2012), most recent reports suggest that globaldamages caused by cybercrime will surpass $8 trillion in 2023 and $10.5 trillion by 2025 (Cybersecurity Ventures 2022). To put it in perspective, only eighteen countries in the world had a GDP in 2022 larger than one trillion dollars.

ling for a multitude of firm-level characteristics. We also find that firms' profitability outcomes do not change with cyberrisk because the risk is hedged by innovation. In other words, cybercrime can drive innovation by forcing companies to improve their security measures and systems, which can lead to new technology and products being developed. The need to protect against cyber threats creates a demand for more secure software, hardware, and services, which can lead to technological advancements and to sustained long-term growth.

Our project contributes to multiple strands of literature. First, we contribute to a recent literature on data as a main driver of economic growth. We extend the theoretical framework in Farboodi et al. (2019) and Farboodi and Veldkamp (2021) to include cybercrime in an endogenous growth model where data is a key input for prediction. In this literature, data is a valuable asset that helps firms reduce uncertainty about some optimal production technique, making them approach some optimal benchmark. In our framework, data has another crucial economic role besides prediction, namely as an input in the production of ideas. The role of data for innovation, highlighted first by Jones and Tonetti (2020), is key to justify the presence of endogenous growth in our setup. In our framework, data is both a valuable information asset used for prediction but also a technology associated with the production of ideas. Thus, while data can be stolen and damaged, leading to lower aggregate output and knowledge, it can also be a vehicle for sustained growth when it is a driver of expanding varieties or when it expands the innovation possibility frontier.

The way in which data is modeled has non-trivial implications for long-run growth. If data is used as an input to expand the innovation frontier, as we consider in this paper, it can be a vehicle for sustained growth; in opposition, if the role of data is to be used only in prediction and, therefore, only to reduce uncertainty, there is a lower bound that cannot be overcome (uncertainty cannot fall below zero), and therefore data, by itself, cannot be a force conducting to sustained long- term growth. Other growth models, such as Hou et al. (2022), constrain growth by bounding the economy's data storage capacity. Although data may contribute to unlimited growth, data storage is, in itself, a limit to growth. In our framework, growth is both encouraged and bounded by the presence of cybercrime.

Other studies on the interaction between data and growth examine the trade-off between trading data with third-parties and privacy concerns. Cong et al. (2022) and Cong et al. (2021) develop endogenous growth models with consumer-generated data as a new factor for knowledge accumulation. Canayaz et al. (2022) develop a

model where data privacy laws limit the acquisition, processing, and trade of consumer personal data in order to examine the heterogeneous affects on firms with and without previously gathered customer data.

We also contribute to the literature on the consequences of cybercrime on firm financial and economic outcomes. Florackis et al. (2023) develops a measure of corporate cyberrisk for the period 2007-2018 for approximately 3100 U.S.-based publicly-listed firms and finds that this risk is priced in the stock market in the form of higher future returns. Kamiya et al. (2021) studies the financial performance of firms that are successfully cyber-attacked, as well as the ex-ante characteristics of those firms that are attacked. Both these studies show that cyberrisk is ex-ante positively correlated with firm size, growth opportunities (Tobin's Q), profitability (ROA) and expenditures of research and development (R&D), but R&D expenses are not correlated with the ex-post probability of a cyberattack. Moreover, those firms that are successfully attacked experience negative cumulative abnormal returns around the attack, and attacks have a significant negative long-term impact on sales growth, customer confidence, and in operating performance. Relative to these studies, we examine the impact of cybercrime on firms' innovation activities.

We use the measure of cyberrisk developed in Florackis et al. (2023) to investigate whether companies that are highly exposed to cyberrisk hedge themselves by innovating more. The novelty of our approach is to examine the endogenous response of firms subject to heterogeneous levels of cyberrisk; and we find that firms do mitigate the negative effects of cyberrisk by innovating more. Moreover, we examine multiple measures of innovation, such as patent counts, patent varieties, patenting times, but also firm boundaries and firm trademarks.

The rest of the paper proceeds as follows. Section 2 presents our theoretical framework: we first discuss a simple data economy without cybercrime in Subsection 2.1. We then add cybercrime and cybersecurity in Subsection 2.2 and compare and contrast the outcomes. We present our theory of cybercrime driven innovation in Subsection 2.3. Section 3 presents our empirical analysis and results. Lastly, Section 4 concludes.

# 2 Theoretical framework

## 2.1 A data economy without cyber-crime

In this section, we briefly present a growth-data model based on Farboodi et al. (2019) and Farboodi and Veldkamp (2021) to set the stage for the subsequent analysis and discussion of the growth implications of cybercrime. In the model, data is a by-product of economic activity and helps firms reduce the uncertainty in their optimal production technique. Also as in the original model, producers acquire data from customers, as a by-product of market transactions, and they can also trade data with one another. Differently from the benchmark model, our setup abstracts from capital accumulation (each firm is endowed with a fixed unit of capital), and it considers two distinct groups of producers which will be designated, further below, as high data-intensity firms and low data-intensity firms.

The economy is populated by a large number of firms, indexed by $i$, which produce different varieties of a final good. At date $t$, a given producer $i$ generates a variety of quality $A_{i,t}$. Because the single input employed in production is one unit of capital, variable $A_{i,t}$ also represents the real value of the producer's output. On the aggregate, the economy's total income amounts to:

$$Y_t = \int_i A_{i,t} di \tag{1}$$

The quality of variety $i$ is determined by the employed production technique $a_{i,t}$. If $a_{i,t}$ corresponds to the optimal technique (i.e., to the most productive technique under the current state of technology), a maximum quality level $\overline{A}$ is accomplished. However, the optimal technique is not known with certainty; it is a stochastic variable with two components, one persistent ($\theta_t$) and another one transitory ($\epsilon_{a,i,t}$). The transitory component might be interpreted as an unlearnable quality shock, $\epsilon_{a,t} \sim i.i.d.(0, \sigma_a^2)$; the persistent component is modelled under the form of an order 1 autoregressive process,

$$\theta_t = \overline{\theta} + \rho(\theta_{t-1} - \overline{\theta}) + \eta_t \tag{2}$$

with $\overline{\theta} > 0$, $\rho \in (0, 1)$, and $\eta_t \sim i.i.d.(0, \sigma_\theta^2)$. Variable $\eta_t$ is designated as fundamental uncertainty.

The quality of the variety produced by firm $i$ is defined as the difference between maximum quality and the squared distance between the selected production technique and the optimal technique, i.e.,

$$A_{i,t} = \overline{A} - [a_{i,t} - (\theta_t + \epsilon_{a,i,t})]^2 \tag{3}$$

In this setting, data serves the purpose of lowering uncertainty. Because producers cannot learn about $\epsilon_{a,i,t}$, data is essentially informative about $\theta_t$, i.e., producers employ data to infer $\theta_t$ and to choose a technique of production as close as possible to the maximum quality level. Producers collect data from clients in an amount $z_i$; this value is a by-product of economic activity, and it is contingent on the ability to mine data from customers (in the current setting, this value is assumed constant over time). Each data point collected by the producer reveals information on $\theta_t$ with a given degree of accuracy. Signal noise, for each data point $m$, amounts to: $\epsilon_{m,i,t} \sim i.i.d.(0, \sigma_\epsilon^2)$. Notice that there are three sources of uncertainty associated with production: producers face unlearnable shocks, fundamental uncertainty, and signal noise. These are all relevant in shaping agents' decisions and the steady state levels of the endogenous variables in the model.

Later on in the analysis, when characterizing equilibrium conditions, one will realize that the existence of a steady state requires the volatility of the unlearnable shock not to be excessively large relative to the volatility attached with the fundamental uncertainty component. Specifically, the following constraint suffices to guarantee the existence of the steady state and, therefore, from this point forward, the constraint is imposed to the model,

$$\sigma_a^2 \leq 4\sigma_\theta^2 \tag{4}$$

As in Farboodi and Veldkamp (2021), it is assumed that data can be traded across production units. Variable $\delta_{i,t}$ represents the amount of data traded by firm $i$; this can be a positive value (if or when the firm is acquiring data) or a negative value (if or when the firm is selling data). When the producer buys data in an amount $\delta_{i,t} > 0$, its available stock of data becomes $z_i + \delta_{i,t}$. If the producer sells data in an amount $\delta_{i,t} < 0$, then the data stock falls to $z_i + \iota\delta_{i,t}$, with $\iota \in (0, 1)$. Parameter $\iota$ reflects the

partial non-rivalry of data, i.e., as the firm sells data, it can continue to partially make use of it ($\iota$ is the fraction of lost data through selling). Defining $\Delta_{i,t} \equiv \mathbf{1}_{\delta_{i,t}>0} + \iota \mathbf{1}_{\delta_{i,t}<0}$, the stock of data available for any firm $i$ to develop its activity at date $t$, will then be $z_i + \Delta_{i,t}\delta_{i,t}$. The price at which data is traded across producers is represented by variable $\pi_t$.

The information set of production unit $i$ contemplates information about production techniques and forecasting signals. This information set, $\Im_{i,t}$, is relevant for computing the producer's stock of knowledge, $\Omega_{i,t} \geq 0$. If one interprets the stock of knowledge as corresponding to the precision of the forecast about $\theta_t$ (i.e., the inverse of the respective variance), it can be expressed under the form

$$\Omega_{i,t} \equiv \mathbf{E}_i \left[ \left( \mathbf{E}_i \left[ \theta_t | \Im_{i,t} \right] - \theta_t \right)^2 \right]^{-1} \tag{5}$$

Given the definition of the stock of knowledge (5), the quality of the produced good, presented in equation (3), will be equivalent to,

$$A_{i,t} = \overline{A} - \left( \Omega_{i,t}^{-1} + \sigma_a^2 \right) \tag{6}$$

In equation (6), $\Omega_{i,t}^{-1} + \sigma_a^2$ denotes the conditional variance of the optimal technique. Observe that, given the presence of unlearnable uncertainty, the maximum quality level $\overline{A}$ will never be reached, independently of how large the stock of knowledge is. In fact, if $\Omega_{i,t} \to +\infty$ then $A_{i,t} \to \overline{A} - \sigma_a^2$.

The problem faced by firm $i$ consists in maximizing the expected value of its flow of present and future profits. Revenues correspond to its output, as presented in equation (6), while costs are those associated with the traded data and with the acquisition of the unit of capital required for production (designate this latter cost by $r > 0$). The objective function of the producer is, thus,

$$V_{i,0} \equiv \mathbf{E}_0 \sum_{t=0}^{\infty} \beta^t \left( A_{i,t} - \pi_t \delta_{i,t} - r \right) \tag{7}$$

In equation (7), $\beta \in (0,1)$ represents the intertemporal discount factor. The maximization of (7) is subject to a constraint on the motion of the stock of knowledge. This constraint takes the form of a difference equation,

$$\Omega_{i,t+1} = \left[\rho^2(\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + (z_i + \Delta_{i,t}\delta_{i,t})\,\sigma_\epsilon^{-2} \tag{8}$$

In equation (8), the second term on the r.h.s. represents the inflows and outflows of data (with $\sigma_\epsilon^{-2}$ indicating the additional information learned at each period), while the first term reflects the impact of uncertainty over the stock of knowledge (specifically, observe that if $\sigma_a^2 = \sigma_\theta^2 = 0$ then $\Omega_{i,t+1} \to \infty$; and if $\sigma_a^2, \sigma_\theta^2 \to \infty$ then $\Omega_{i,t+1}$ is exclusively determined by the second term of the r.h.s. of the equation; observe, as well, for fixed uncertainty levels, that the larger is the current stock of knowledge, $\Omega_{i,t}$, the larger will also be the future stock of knowledge, $\Omega_{i,t+1}$, given the process of knowledge accumulation).

The optimization problem of firm $i$ corresponds to the maximization of $V_{i,0}$, as presented in equation (7), subject to the constraint on the evolution of the stock of knowledge, displayed in (8). This is an optimal control problem with two endogenous variables: a state variable, $\Omega_{i,t}$, and a control variable, $\delta_{i,t}$ (the firm chooses how much data to sell or buy with the objective of maximizing profits). Variable $z_i$ is exogenous and constant, and $\pi_t$ is endogenously determined, although its derivation becomes possible only after taking an aggregate perspective over the economy (further below). All the other elements in the model are interpreted as parameters, including the three relevant variances.

### 2.1.1 Heterogeneous producers

In the proposed data-growth model, the underlying economic environment encompassed a large number of firms, identical to one another in every respect: all firms maximize profits, given a constraint on the evolution of the corresponding stock of knowledge. A straightforward corollary of the mimetic behavior of firms is that data trading will not take place, because the optimal strategy would be the same for all producers: they all would want to sell / buy data, but they would have no other firm with whom to trade. Data trading requires firm heterogeneity, and this is now introduced by splitting the universe of production units into two homogeneous groups: the high data-intensity producers (indexed by $h$) and the low data-intensity producers (indexed by $l$). Producers in group $h$ ($l$) correspond to a constant share $u$ ($1-u$) of the universe firms.

The distinguishing features between the two types of firms are assumed to be the following:

1) High data-intensity firms have the ability and the opportunity to extract or mine a larger amount of data from the interaction with customers than the firms in the low data-intensity category: $z_h > z_l$ (producers in group $h$ take data mining as a primordial part of their activity, much more than producers in group $l$ do).

2) Low data-intensity producers have access to other sources of knowledge, besides data-driven knowledge, to develop their activity (e.g., accumulated practical experience). High data-intensity producers do not have access to such knowledge.

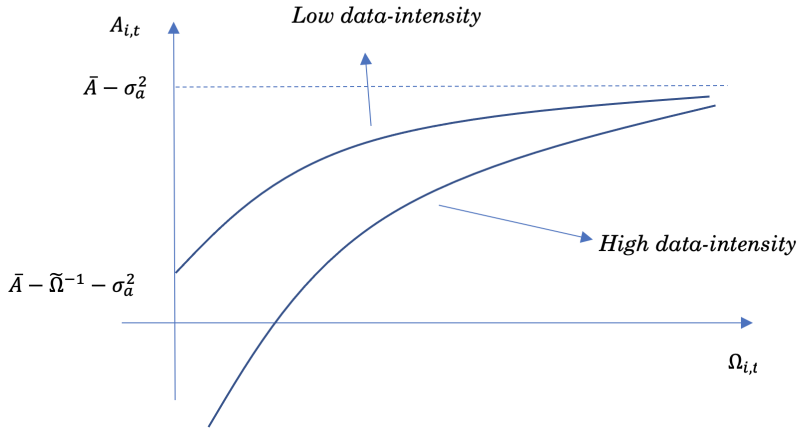
The second assumption in the above list implies that the only relevant knowledge accessible to a high data-intensity firm to improve the quality of its good's variety is associated with digital data; hence, for firms in group $h$, the corresponding output is defined exactly as in equation (6). For firms in group $l$, there exists an additional stock of knowledge, of a non-digital nature, which is modelled as a constant value $\widetilde{\Omega} > 0$. Thus, for low data-intensity producers, the quality of the generated good's variety is, rather than (6),

$$A_{i,t} = \overline{A} - \left[ \left( \Omega_{i,t} + \widetilde{\Omega} \right)^{-1} + \sigma_a^2 \right], \quad i = l \tag{9}$$

This distinguishing feature between the two types of producers is depicted in Fig. 1. High data-intensity production units require more data to approach the quality ceiling than the low data-intensity firms. The difference becomes smaller as the data-based stock of knowledge increases.

Note, as displayed in Fig. 1, that quality ceilings are taken as being identical across firms in the two groups. This assumption is adopted with the objective of maintaining some similarity across firms, thus keeping the heterogeneity circumscribed to the two aforementioned features. These features enclose two countervailing forces. On one hand, high data-intensity firms are able to collect an amount of data larger than the quantity of data extracted by low data-intensity firms. On the other hand, the requirements of data by producers in the $h$ category are larger than the ones of producers of type $l$ (which also use other, non-digital, sources of data). Hence, high data-intensity firms access more data but also need more data to optimize production, what makes it hard to discern, at first sight, which group will be the one profiting from selling data and which group will be the one profiting from buying data. To reach a result, the optimal control

Figure 1: Big data stock of knowledge and good's quality in low data-intensity and high data-intensity sectors.




Legend: The X-axis depicts the stock of knowledge. The Y axis depicts the product quality. High data-intensity production units require more data to approach the quality ceiling than the low data-intensity firms. The difference becomes smaller as the data-based stock of knowledge increases.

problem of the representative firm in each of the two groups must be approached. As designed, the model allows both possibilities to be simultaneously optimal, for different data transaction prices.

### 2.1.2 Dynamics and the steady state

Take some producer $i = h, l$, and write the current-value Hamiltonian function associated with the intertemporal maximization problem,

$$\begin{aligned} H(\Omega_{i,t}; \delta_{i,t}; p_{i,t}) &= (A_{i,t} - \pi_t \delta_{i,t} - r) \\ &+\beta p_{i,t+1} \left\{ \left[\rho^2 (\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + (z_i + \Delta_{i,t}\delta_{i,t})\, \sigma_\epsilon^{-2} - \Omega_{i,t} \right\} \end{aligned} \tag{10}$$

In expression (10), $p_{i,t}$ is the co-state variable attached to the stock of knowledge. First-order optimality conditions are:

$$\frac{\partial H}{\partial \delta_{i,t}} = 0 \Longrightarrow \beta p_{i,t+1} = \frac{\pi_t \sigma_\epsilon^2}{\Delta_{i,t}} \tag{11}$$

and,

$$\beta p_{i,t+1} - p_{i,t} = -\frac{\partial H}{\partial \Omega_{i,t}} \Longrightarrow \left[\rho + \frac{\sigma_\theta^2}{\rho}(\Omega_{i,t} + \sigma_a^{-2})\right]^{-2} \beta p_{i,t+1} = p_{i,t} - \left(\Omega_{i,t} + \Omega_{(l,h)}\right)^{-2} \quad (12)$$

with $\Omega_{(l,h)} = 0$ for firms in group $h$ and $\Omega_{(l,h)} = \widetilde{\Omega}$ for firms in group $l$. The transversality condition $\lim_{t\to\infty} \Omega_{i,t}\beta^t p_{i,t} = 0$ must be satisfied as well.

Combining optimality conditions (11) and (12), one obtains the following equality,

$$\left[\rho + \frac{\sigma_\theta^2}{\rho}(\Omega_{i,t} + \sigma_a^{-2})\right]^{-2} \frac{\pi_t \sigma_\epsilon^2}{\Delta_{i,t}} = \frac{\pi_{t-1}\sigma_\epsilon^2}{\beta \Delta_{i,t-1}} - \left(\Omega_{i,t} + \Omega_{(l,h)}\right)^{-2} \quad (13)$$

Next, we define the economy's steady state.

**Definition 1** *The economy reaches a steady state when the following conditions are met: (i) the position of a firm as buyer or seller of data remains unchanged,* $\Delta_i^* \equiv \Delta_{i,t} = \Delta_{i,t-1}$*; (ii) data traded by each firm is a constant value,* $\delta_i^* \equiv \delta_{i,t} = \delta_{i,t-1}$*; (iii) the data-trading price is constant,* $\pi^* \equiv \pi_t = \pi_{t-1}$*.*

Under the above conditions, the steady state stocks of knowledge, if they exist, also correspond to constant values, $\Omega_i^*$. To characterize the equilibrium, recall that there are two types of firms, $h$ and $l$, which are homogeneous groups, and therefore data trading can only occur across groups and not within groups. Hence, the analysis of the steady state requires distinguishing between two alternative scenarios:

*a*) High data-intensity producers are data sellers and low data-intensity producers are data buyers. In this case, the steady state version of (13) is, for firms in the $h$ group,

$$\left[\rho + \frac{\sigma_\theta^2}{\rho}(\Omega_h^* + \sigma_a^{-2})\right]^{-2} \frac{\pi^* \sigma_\epsilon^2}{\iota} = \frac{\pi^* \sigma_\epsilon^2}{\beta \iota} - (\Omega_h^*)^{-2} \quad (14)$$

and, for firms in the $l$ group:

$$\left[\rho + \frac{\sigma_\theta^2}{\rho}(\Omega_l^* + \sigma_a^{-2})\right]^{-2} \pi^* \sigma_\epsilon^2 = \frac{\pi^* \sigma_\epsilon^2}{\beta} - \left(\Omega_l^* + \widetilde{\Omega}\right)^{-2} \quad (15)$$

*b*) High data-intensity producers are data buyers and low data-intensity producers are data sellers. In this case, the steady state version of (13) is, for firms in the $h$ group,

$$\left[\rho + \frac{\sigma_\theta^2}{\rho}(\Omega_h^* + \sigma_a^{-2})\right]^{-2} \pi^* \sigma_\epsilon^2 = \frac{\pi^* \sigma_\epsilon^2}{\beta} - (\Omega_h^*)^{-2} \tag{16}$$

and, for firms in the $l$ group:

$$\left[\rho + \frac{\sigma_\theta^2}{\rho}(\Omega_l^* + \sigma_a^{-2})\right]^{-2} \frac{\pi^* \sigma_\epsilon^2}{\iota} = \frac{\pi^* \sigma_\epsilon^2}{\beta \iota} - \left(\Omega_l^* + \widetilde{\Omega}\right)^{-2} \tag{17}$$

In the steady state, it is possible to derive, from the constraint equation (8), the value of the amount of data traded by each firm,

$$\delta_i^* = \frac{\left\{\Omega_i^* - [\rho^2(\Omega_i^* + \sigma_a^{-2})^{-1} + \sigma_\theta^2]^{-1}\right\} \sigma_\epsilon^2 - z_i}{\Delta_i^*} \tag{18}$$

Expression (18) applies to all four cases depicted above, although one should note that $\delta_i^* > 0$ if firm $i$ is a data buyer ($\Delta_i^* = 1$) and $\delta_i^* < 0$ if producer $i$ is a data seller ($\Delta_i^* = \iota$). Observe the relevance of variable $z_i$ in equation (18): the relative position of firms regarding the acquisition and the selling of data is in a large extent dependent on the amount of data each of the types of firms can extract from their customers.

On the aggregate, data-selling and data-purchasing cancel one another and, thus, $\int_i \delta_i^* di = 0$. Because only two types of identical firms exist, and the high data-intensity producers are a percentage $u$ of the universe of firms, the previous condition is equivalent to: $u\delta_h^* + (1-u)\delta_l^* = 0$. Terms $|u\delta_h^*|$ and $|(1-u)\delta_l^*|$ both represent the amount of traded data (data sold by one group of agents and bought by the other group).

Given the output aggregator (1), the steady state total income of the economy can be represented under the form (with the number of firms normalized to 1),

$$Y^* = \overline{A} - \left[u\left(\Omega_h^*\right)^{-1} + (1-u)\left(\Omega_l^* + \widetilde{\Omega}\right)^{-1} + \sigma_a^2\right] \tag{19}$$

The characterization of the steady state of the model allowed to uncover the following relevant outcomes:

(*i*) Similar to Farboodi and Veldkamp (2021), data by itself is not a generator of sustained endogenous growth. Data increases the stock of knowledge that will reduce uncertainty and, thus, assist the producer in choosing a production technique closer to the optimum. However, as it approaches the optimum, the economy loses space to continue growing.

(*ii*) The zero-growth steady state is translated into constant steady state stocks of knowledge. These stocks of knowledge are distinct across types of producers.

(*iii*) Depending on the data-mining capacity and other features, both types of firms - low data-intensity and high data-intensity - can be buyers or sellers of data. When one of the groups is a seller, the other is a buyer (i.e., data is traded across groups of firms).

(*iv*) The amounts of data sold and bought in the steady state are constant, and the price at which transactions occur, in such scenario, is also constant.

(*v*) Because the stocks of knowledge of each producer are constant in the steady state, the steady state level of aggregate output is constant as well. This is, in fact, for now, a model that can be associated with neoclassical growth: the accumulation of an input (in this case, data) conducts the economy to a long-term scenario of zero growth (diminishing marginal returns prevail).

### 2.1.3 Data Demand and Data Supply

The steady state analysis of the previous section left various questions unanswered: in which conditions one or the other type of firms is a buyer or a seller of data? How are the several steady state values - stock of knowledge, price of data, stocks of traded data - related with one another? Does the equilibrium exist in every circumstance and is it unique? To answer these interrogations, let us approach the steady state from the perspective of data demand and data supply.

On the demand side (i.e., from the perspective of data buyers), equations (15) and (16) establish relations between the price of data and the stocks of knowledge of the firms (in each of the two assumed trading scenarios). As it is straightforward to observe, these relations are of opposite sign (in the steady state, a higher trading price is synonymous of lower stocks of knowledge). The demanded quantity of data is $q^d = (1-u)\delta_l^*$ or $q^d = u\delta_h^*$, depending on whether the data-purchasers are firms in group $l$ or firms in group $h$. Demanded quantities are also attached to the stocks of knowledge, via (18), and this relation is, under constraint (4), a relation of identical

sign.[2] Therefore, there exists a function $\pi^* = f(\Omega_i^*)$, $f' < 0$, and another function $q^d = f(\Omega_i^*)$, $f' > 0$; together they imply a demand curve $q^d = f(\pi^*)$, $f' < 0$. Each point on the demand curve is obtained for some constant value $\Omega_i$, $\forall i \in \{l, h\}$.

On the supply side, the reasoning is similar. Equations (14) and (17) are such that $\pi^* = f(\Omega_i^*)$, $f' < 0$. Supplied quantities of data are, in each of the two assumed scenarios, $q^s = -u\delta_h^*$ and $q^s = -(1-u)\delta_l^*$; in both cases, $q^s = f(\Omega_i^*)$, $f' < 0$. Therefore, the supply curve is such that $q^s = f(\pi^*)$, $f' > 0$. In the intersection of the two curves (supply and demand) one finds the steady state equilibrium values of $\pi^*$ and $q^*$ [$q^* = (1-u)\delta_l^* = -u\delta_h^*$ or $q^* = u\delta_h^* = -(1-u)\delta_l^*$, depending on whether firms in group $h$ are data suppliers or data purchasers, respectively]. To determine $\Omega_h^*$ and $\Omega_l^*$, one may solve the price equations for the obtained equilibrium price.

Constraint (4), besides guaranteeing a positive slope for the supply curve and a negative slope for the demand curve, also assures that the equilibrium price is positive.[3] Hence, with a negatively sloped demand curve and a positively sloped supply curve that intersect at some positive price, one guarantees that the equilibrium exists and is unique.[4] To illustrate the equilibrium outcome, take a numerical example. Consider Table 1. The table contains a set of benchmark parameter values.

---

[2] To confirm this assertion, compute $\frac{\partial q^d}{\partial \Omega_i^*}$. The derivative is a positive value if $\left[\rho + \frac{\sigma_\theta^2}{\rho}(\Omega_i^* + \sigma_a^{-2})\right]^2 > 1$, what is equivalent to $\Omega_i^* > \frac{\rho(1-\rho)}{\sigma_\theta^2} - \sigma_a^{-2}$. Under constraint (4), this is a true condition, and therefore a relation of opposite sign is established between the stock of knowledge and demanded data.

[3] For any of the steady state equations involving the price, i.e., (14) to (17), a positive price is guaranteed under condition $\left[\rho + \frac{\sigma_\theta^2}{\rho}(\Omega_i^* + \sigma_a^{-2})\right]^2 > \beta$. Given constraint (4), this is a true condition for any admissible parameter values.

[4] Although the equilibrium exists for any admissible parameter values, including the original data endowments (i.e., the data mined from customers), the data endowments may impose an equilibrium where only one of the two trading possibilities is feasible. This point is discussed further below.

| **Parameter / exogenous variable** | **Symbol** | **Value** |
|---|---|---|
| Coefficient of the AR(1) process | $\rho$ | 0.9 |
| Maximum quality | $\overline{A}$ | 1 |
| Variance of the fundamental uncertainty | $\sigma_\theta^2$ | 0.25 |
| Variance of the signal noise | $\sigma_\epsilon^2$ | 0.25 |
| Variance of the unlearnable quality shock | $\sigma_a^2$ | 0.25 |
| Share of data lost when production units sell data | $\iota$ | 0.6 |
| Intertemporal discount factor | $\beta$ | 0.96 |
| Collected data (low data-intensity) | $z_l$ | 0.75 |
| Collected data (high data-intensity) | $z_h$ | 1.5 |
| Share of firms in the high data-intensity sector | $u$ | 0.25 |
| Non-digital data parameter | $\widetilde{\Omega}$ | 2.5 |

Table 1 - Values of parameters and exogenous variables

With the values in Table 1, it is straightforward to obtain steady state results for the endogenous variables. The computed results are systematized in Table 2.

| | $h$ **sellers** / $l$ **buyers** | $h$ **buyers** / $l$ **sellers** |
|---|---|---|
| $\pi^*$ | 0.0448 | 0.0364 |
| $q^*$ | 0.1800 | 0.0750 |
| $\Omega_l^*$ | 7.0537 | 5.7657 |
| $\Omega_h^*$ | 7.3894 | 10.4810 |
| $A_l^*$ | 0.6453 | 0.6290 |
| $A_h^*$ | 0.6147 | 0.6546 |
| $Y^*$ | 0.6377 | 0.6354 |

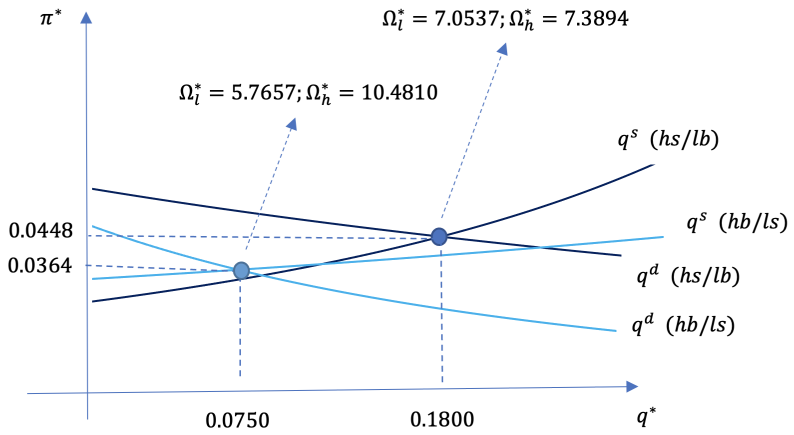
Table 2 - Steady state results (numerical example)

In the proposed numerical example, price and volume of traded data are higher in the $h$ sellers - $l$ buyers scenario than in the opposite case. High data-intensity firms accumulate more big data knowledge than low data-intensity firms, in any of the cases; however, one should keep in mind that firms in group $l$ also access non-digital data, what justifies the relatively higher output level in the $h$ sellers - $l$ buyers scenario. Given formula (19), the value of aggregate output is computed as well; this value does not

differ significantly across the two data trading alternatives. Recall that the value of output has a ceiling, such that $Y^* \leq \overline{A} - \sigma_a^2 = 0.75$. The main conclusion emerging from the simultaneous analysis of the two cases is that any trading solution can be an optimal result: firms in each of the groups may become buyers or sellers of data, for the same parameter values characterizing the state of the economy. The differences between the two scenarios are the price at which data is exchanged among producers and the traded volume. This has impact on the accumulated knowledge and on the ability to generate output.

Fig. 2 displays the demand and supply schedule for this numerical example, considering both alternatives regarding the exchange of data. The equilibrium points correspond to the loci in which the transaction of data generates steady state levels of knowledge that serve to choose the best attainable technique of production.

Figure 2: Data demand, data supply, and the steady state equilibrium.



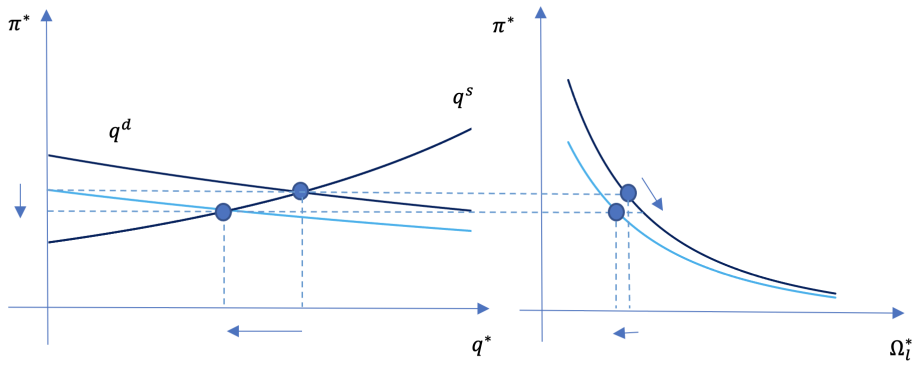

Legend: The model features two types of equilibria, depending on whether high data intensity firms are sellers and low data intensity firms are buyers (in dark-blue), in which case the price of data is 0.0448 and the quantity of data is 0.1800, or whether high data intensity firms are buyers and low data intensity firms are sellers (in light blue), in which case the price of data is 0.0364 and the quantity of data is 0.0750.

The demand and supply schedule is useful to inquire about how the steady state is disturbed when there are perturbations in the values of relevant parameters. Consider, sequentially, four positive changes, on the values of parameters $\widetilde{\Omega}$, $u$, $z_h$ and $z_l$.

In the first case, an increase in the stock of non-data knowledge $\widetilde{\Omega}$, held by firms in group $l$, will move the demand curve down, if firms in this group are data buyers,

and move the supply curve down if these firms are data sellers. As a result, in both cases, the steady state price of data declines, while the quantity of exchanged data will decrease (if $l$ firms are data buyers) or increase (if $l$ firms are data sellers). Because the price of data falls, the steady state stock of knowledge of producers in group $h$, $\Omega_h^*$, increases [there is a movement to the right along the $(\Omega_h^*, \pi^*)$ schedule]; the impact on $\Omega_l^*$ is not as straightforward to identify because besides the downward movement along the $(\Omega_l^*, \pi^*)$ curve, triggered by $\Delta\widetilde{\Omega} > 0$, this perturbation also shifts the $(\Omega_l^*, \pi^*)$ curve to the left (the positive change in $\widetilde{\Omega}$ makes $\pi^*$ fall for each potential value of $\Omega_l^*$), provoking an eventual decline on the equilibrium level of $\Omega_l^*$. Fig. 3 depicts this case, by assuming that firms in group $l$ are data buyers. Two graphics are shown; the one in the left represents the demand-supply schedule and the corresponding equilibrium perturbation, while the graphic on the right side displays the effect over $\Omega_l^*$.

Figure 3: Steady state perturbation: $\Delta\widetilde{\Omega} > 0$

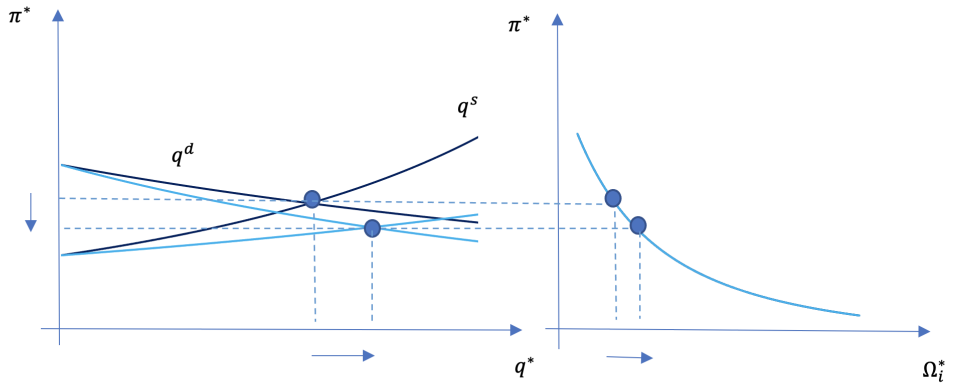



Legend: The left hand-side figure shows the demand-supply schedule and the corresponding equilibrium perturbation, while the graphic on the right side displays the effect over $\Omega_l^*$.

Our second comparative statics exercise involves the shares of producers in each sector. Assume that there is a positive change in the fraction of firms in the $h$ sector, $\Delta u > 0$. When high data-intensity producers are data suppliers and low data-intensity producers are data purchasers, the perturbation in $u$ triggers a downward shift in both the demand and the supply curves [because the demanded and the supplied quantities are, respectively, $(1-u)\delta_l^*$ and $-u\delta_h^*$]; in the opposite case, for $h$ buyers and $l$ sellers, both curves will shift upward [because the demanded and the supplied quan-

tities are, respectively, $u\delta_h^*$ and $-(1-u)\delta_l^*]$. Therefore, the change in the price and in the exchanged quantities of data originating on a change on the composition of the industries will be contingent on whether each group buys or sells data. To illustrate this case graphically concentrate in the $h$ sellers / $l$ buyers scenario; in this scenario, the equilibrium price decreases and the exchanged amount of data increases, relative to the benchmark case. Note, as well, that $u$ does not directly influence the relation between $\pi^*$ and $\Omega_i^*$, meaning that the schedule $(\Omega_i^*, \pi^*)$ will not move $\forall i \in \{l, h\}$.; therefore, the fall in the equilibrium price will implicate a downward movement along the curve $(\Omega_i^*, \pi^*)$, thus leading to an increase in the data-driven stocks of knowledge. Fig. 4 proceeds with the respective graphical representation. Observe that the stocks of knowledge increase in the circumstance in which $h$ are sellers and $l$ are buyers, but they would decrease if $h$ were data buyers and $l$ were data sellers (due to the increase in the equilibrium price).

Figure 4: Steady state perturbation: $\Delta u > 0$




Legend: The left hand-side figure shows the demand-supply schedule and the corresponding equilibrium perturbation, while the graphic on the right side displays the effect over $\Omega_l^*$.

Next, take the case $\Delta z_h > 0$. A change in the data endowment of high data-intensity producers exerts influence on the supply curve if firms $h$ are data sellers and on the demand curve if firms $h$ are data buyers [given (18)]. In the respective cases, each of the curves moves down, in the direction of the horizontal axis, given the positive change in the data endowment. Regardless of the trading position of each group of firms, the data transaction price falls. Because data endowments do not directly influence the

steady state relation between equilibrium price and steady state stocks of knowledge, the perturbation will provoke a movement along the curve $(\Omega_i^*, \pi^*)$, such that the stocks of knowledge (of both types of producers) will suffer positive changes. As expected, when additional data is extracted from customers (in this case, by high-data intensity production units), the stocks of knowledge will increase.

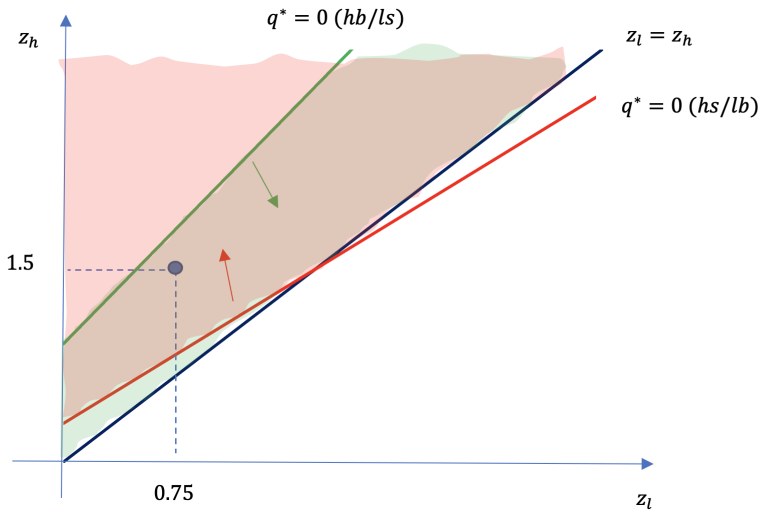
The stocks of knowledge of every producer also increase for $\Delta z_l > 0$. To confirm that this is a true assertion, observe that in this case the demand curve is disturbed (shifts downward) when low data-intensity firms are data buyers, and the supply curve moves (also shifts downward) when low data-intensity firms are data sellers [for the relation in (18)]. Again, the equilibrium data-trading price declines, triggering, given equations (14) to (17), an increase in the stocks of knowledge. Therefore, if the producers in group $l$ extract additional data from customers, then the price of data will decline, the stocks of knowledge of both groups of firms will increase, and their output will increase accordingly. Obviously, if $\Delta z_h < 0$ or $\Delta z_l < 0$ then the opposite effect would take place: the data transaction price would increase, and stocks of knowledge and quality of produced goods would fall.

One relevant question is whether trade is feasible for every possible data endowments (i.e., for any $z_l > 0$ and $z_h > z_l$). There are constraints on the values that these endowments can take, which are imposed by equation (18). In particular, relatively low levels of data availability when firms in group $h$ assume the selling position or relatively high levels of data availability when firms in group $h$ assume the buying position, may turn data trading unfeasible. Recovering the parameter values in Table 1, it is possible to graphically represent the trading regions in the $(z_l, z_h)$ space. Fig. 5 displays three lines in the $(z_l, z_h)$ space: the boundary $z_h = z_l$, imposed as an assumption, and the zero-trading boundary $q^* = 0$, for each of the scenarios ($h$ sellers / $l$ buyers and $h$ buyers / $l$ sellers).

When firms in the $h$ group supply data and firms in the $l$ group demand data, trade is possible in the region above $q^* = 0$, as long as $z_h > z_l$. In this case, as we depart from the frontier in the direction of larger data endowments, the larger will be the quantity of traded data. When firms in the $h$ group demand data and producers in group $l$ supply data, trade is feasible in the region below $q^* = 0$ (and above $z_h = z_l$). In this second case, as we depart from the frontier in the direction of smaller data endowments, the larger will be the quantity of traded data. Looking at the two possibilities together, one realizes that there is a a region of data endowments for which each class of producers may assume any of the two trading positions; outside such region, data buying or data

selling are possible for any of the groups, but not the two options simultaneously. In the graphic, the point in the example ($z_l = 0.75; z_h = 1.5$) is highlighted (and one observes that it lies inside the intersection region).

Figure 5: Trading areas in the $(z_l, z_h)$ space




Legend: Fig.5 displays three lines in the $(z_l, z_h)$ space: the boundary $z_h = z_l$, imposed as an assumption, and the zero-trading boundary $q^* = 0$, for each of the scenarios ($h$ sellers / $l$ buyers and $h$ buyers / $l$ sellers). There is a a region of data endowments for which each class of producers may assume any of the two trading positions; outside such region, data buying or data selling are possible for any of the groups, but not the two options simultaneously. In the graphic, the point in the example ($z_l = 0.75; z_h = 1.5$) is highlighted (and one observes that it lies inside the intersection region).

## 2.2 A data economy with cyber crime and cyber security

Cybercrime may take various forms. A straightforward way to interpret it is to consider that criminal activity provokes a direct and immediate loss of the data the firm holds. Cyber attacks take place after the production unit has collected data from customers and after it has eventually traded data with other units. Assuming that cybercrime implies a loss of usable data in a share $\vartheta \in (0, 1)$, the producer's endowment of data becomes $(1-\vartheta)(z_i + \Delta_{i,t}\delta_{i,t})$. In this scenario, the objective function (7) remains the same, but the knowledge dynamic constraint (8) now incorporates the cybercrime feature,

$$\Omega_{i,t+1} = \left[\rho^2(\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + (1-\vartheta)(z_i + \Delta_{i,t}\delta_{i,t})\,\sigma_\epsilon^{-2} \tag{20}$$

In Appendix (*A*.1), we solve a version of the model where cyber-crime affects the entire knowledge stock, that is where cyber-crime can lead not only to a loss of usable data, but also to the loss of some know-how or algorithms to generate knowledge from the data, such that $\Omega_{i,t+1} = (1-\vartheta)\left\{[\rho^2(\Omega_{i,t}+\sigma_a^{-2})^{-1}+\sigma_\theta^2]^{-1}+(z_i+\Delta_{i,t}\delta_{i,t})\,\sigma_\epsilon^{-2}\right\}$ and we show this has no material effects on the equilibrium or the results.

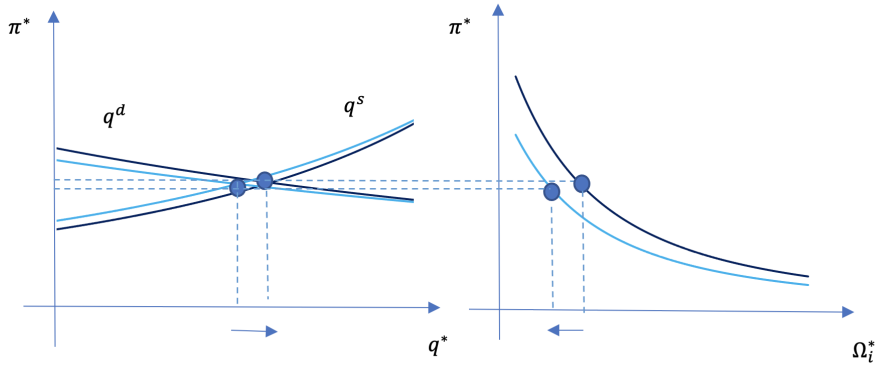
The optimality condition (13) must consider as well this effect of lost data,

$$\left[\rho+\frac{\sigma_\theta^2}{\rho}(\Omega_{i,t}+\sigma_a^{-2})\right]^{-2}\frac{\pi_t\sigma_\epsilon^2}{(1-\vartheta)\Delta_{i,t}}=\frac{\pi_{t-1}\sigma_\epsilon^2}{(1-\vartheta)\beta\Delta_{i,t-1}}-\left(\Omega_{i,t}+\Omega_{(l,h)}\right)^{-2} \tag{21}$$

To investigate the impact of cybercrime over the data market equilibrium and over the steady state stocks of knowledge, one needs, as before, to distinguish between the two trading possibilities. If firms in group $h$ are suppliers of data and firms in group $l$ are purchasers of data, the steady state demand and supply curves will shift to the left with the increase in the value of parameter $\vartheta$. Hence, cyber attacks that damage the access to data will, in this case, contract the amount of traded data, while the price might change (up or down) but not significantly. If firms in group $h$ are buyers of data and firms in group $l$ are sellers of data, cybercrime will move the demand and supply curves to the right, increasing the quantity of traded data and implying an eventual not too significant change in the equilibrium price. From equation (21), one infers that in the steady state there is an opposite sign relation between the price of data and the extent of cybercrime, measured by the value of $\vartheta$. Therefore, in the $(\Omega_i^*, \pi^*)$ schedule, cybercrime will shift the knowledge-price curve closer to the origin. This effect delivers the expected outcome: as the value of parameter $\vartheta$ increases, the stock of knowledge and the quality of output will fall, for firms in any of the two sectors. Fig. 6 displays the data market equilibrium diagram and the knowledge-price curves for the case in which $h$ producers are suppliers of data (the lighter lines are those representing the cybercrime scenario).

Recovering the numerical example of the previous section, steady state results can be quantified. Let $\vartheta = 0.25$ and consider, as well, the array of values in Table 1. Results in Table 3 should be compared with those in Table 2, i.e., with the case without cybercrime.

Figure 6: Steady state equilibrium under cybercrime




Legend: Cybercrime shifts the knowledge-price curve closer to the origin. This effect delivers the expected outcome: as the value of parameter $\vartheta$ increases, the stock of knowledge and the quality of output will fall, for firms in any of the two sectors. This figure displays the data market equilibrium diagram and the knowledge-price curves for the case in which $h$ producers are suppliers of data (the lighter lines are those representing the cybercrime scenario).

| | $h$ **sellers** / $l$ **buyers** | $h$ **buyers** / $l$ **sellers** |
|---|---|---|
| $\pi^*$ | 0.0443 ($-1.12\%$) | 0.0349 ($-4.12\%$) |
| $q^*$ | 0.1519 ($-15.61\%$) | 0.1329 (77.20%) |
| $\Omega_l^*$ | 5.8716 ($-16.76\%$) | 4.8562 ($-15.77\%$) |
| $\Omega_h^*$ | 6.4633 ($-12.53\%$) | 9.2981 ($-11.29\%$) |
| $A_l^*$ | 0.6305 ($-2.29\%$) | 0.6141 ($-2.37\%$) |
| $A_h^*$ | 0.5953 ($-3.16\%$) | 0.6425 ($-1.85\%$) |
| $Y^*$ | 0.6217 ($-2.51\%$) | 0.6212 ($-2.23\%$) |

Table 3 - Steady state results, under cybercrime (numerical example)

In Table 3, the steady state results under cybercrime are presented, as well as the rate of change from the no cybercrime scenario to the scenario with lost or damaged data. The example confirms that the variation in the equilibrium price is small and that traded quantities may increase or decrease, relative to the benchmark setting, depending on the trading position of the firms. Stocks of knowledge, individual output levels, and aggregate output, all suffer a negative change when the availability of data in contracted by cybercrime.

Data subject to cybercrime can be protected. Consider that firms have the faculty of investing in data protection, and that the unitary cost of protection for each data unit that the producer holds is $\widetilde{\tau} > 0$. At point in time $t$, the total expenditure of firm $i$ in cyber-security will be:

$$\tau_{i,t} = \widetilde{\tau}\left(z_i + \Delta_{i,t}\delta_{i,t}\right) \tag{22}$$

It is assumed that incurring in protection costs guarantees partial protection against attacks, i.e., by making investment $\tau_{i,t}$ the data loss effect of the cyberattack falls from share $\vartheta$ to share $\vartheta_\tau \in (0, \vartheta)$. With the protection assumption, each production unit will solve an optimal control problem where the objective function $V_{i,0}$ includes an additional term on the costs side,

$$V_{i,\tau,0} = \mathbf{E}_0 \sum_{t=0}^{\infty} \beta^t \left(A_{i,t} - \pi_t \delta_{i,t} - r - \tau_{i,t}\right) \tag{23}$$

The maximization of (23) is subject to a dynamic constraint similar to (21), but where $\vartheta$ is replaced by $\vartheta_\tau$. The solution of the optimization problem conducts to the following equality,

$$\left[\rho + \frac{\sigma_\theta^2}{\rho}(\Omega_{i,t} + \sigma_a^{-2})\right]^{-2} \frac{\left(\pi_t + \widetilde{\tau}\Delta_{i,t}\right)\sigma_\epsilon^2}{(1-\vartheta_\tau)\Delta_{i,t}} = \frac{\left(\pi_{t-1} + \widetilde{\tau}\Delta_{i,t-1}\right)\sigma_\epsilon^2}{(1-\vartheta_\tau)\beta\Delta_{i,t-1}} - \left(\Omega_{i,t} + \Omega_{(l,h)}\right)^{-2} \tag{24}$$

Equation (24) shares similarities with (21). For both, the introduction of the new terms ($\vartheta$, and $\widetilde{\tau}$ and $\vartheta_\tau$, respectively) pushes the l.h.s. and the r.h.s. of the equilibrium equation upwards. One should expect the extension of the change triggered by the inclusion of parameters $\widetilde{\tau}$ and $\vartheta_\tau$ to have a smaller impact than the extension of the change originating on $\vartheta$; otherwise, the production unit would spend more in cyber-security than the impact it would suffer from cybercrime. Analytically, this condition is, in the steady state,

$$\frac{\left(\pi^* + \widetilde{\tau}\Delta_i^*\right)\sigma_\epsilon^2}{(1-\vartheta_\tau)\Delta_i^*} < \frac{\pi^*\sigma_\epsilon^2}{(1-\vartheta)\Delta_i^*} \Rightarrow \widetilde{\tau} < \frac{\vartheta - \vartheta_\tau}{1-\vartheta}\frac{\pi^*}{\Delta_i^*} \tag{25}$$

Equation (25) reveals that the unitary cost of protection must be a value lower than an expression that involves the extent of cybercrime, the effectiveness of protection, and the cost of trading data; the degree of non-rivalry of data is important as well: data sellers can profitably spend more in cyber-security than data buyers.

Concerning the steady state, the expectable result is such that the stocks of knowledge and the aggregate output level are, under protection, an intermediate result between no crime and crime with no security. Protection mitigates the nefarious impact of cybercrime, thus placing firms in a better position to use their data and create value; however, protection has a direct cost that will certainly hamper the stock of knowledge the firm can accumulate and, therefore, it will reduce the value of output as well. The market equilibrium analysis reveals that cyber-protection shifts both the demand curve and the supply curve down (for both trading positions of the two groups of producers), leading to the formation of a steady state with a lower equilibrium price. As the $(\Omega_i^*, \pi^*)$ schedule moves in the direction of the origin, a new steady state value for $\Omega_i^*, i = l, h$, is formed, which is in fact an intermediate value between those without and with criminal activity. Fig. 7 represents this outcome, by displaying the same diagrams as in previous figures, and where the three circumstances (no crime - crime - protection) are depicted (the graphics are drawn for case $h$ sellers - $l$ buyers).

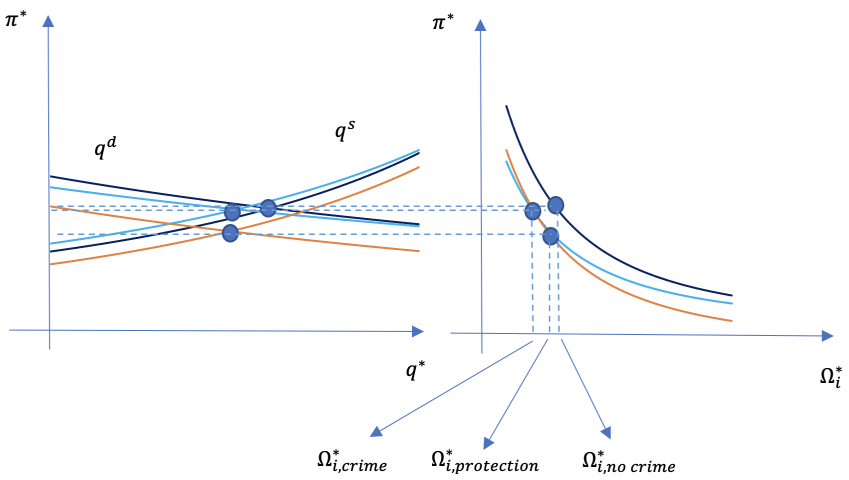
Table 4 indicates the steady state values of the relevant variables of the model, under the implementation of cyber-security measures. These values are all lower than those for the scenario without crime (the percentages indicate the change from the no crime to the security setting).

| | $h$ **sellers** / $l$ **buyers** | $h$ **buyers** / $l$ **sellers** |
|---|---|---|
| $\pi^*$ | 0.0362 ($-17.62\%$) | 0.0292 ($-19.78\%$) |
| $q^*$ | 0.1475 ($-18.06\%$) | 0.0666 ($-11.2\%$) |
| $\Omega_l^*$ | 6.4512 ($-8.54\%$) | 5.4881 ($-4.81\%$) |
| $\Omega_h^*$ | 7.3379 ($-0.70\%$) | 9.6025 ($-8.38\%$) |
| $A_l^*$ | 0.6383 ($-1.08\%$) | 0.6248 ($-0.67\%$) |
| $A_h^*$ | 0.6116 ($-0.50\%$) | 0.6458 ($-1.34\%$) |
| $Y^*$ | 0.6316 ($-0.96\%$) | 0.6301 ($-0.84\%$) |

Legend: The benchmark parameter values are as before. $\widetilde{\tau} = 0.01$ and $\vartheta_\tau = 0.1$.

Table 4 - Steady state results, under cyber-security (numerical example)

Figure 7: Steady state equilibrium with cyber-protection




Legend: With cyber-protection, the stocks of knowledge and the aggregate output level are at an intermediate level between no crime and crime with no protection. Protection mitigates the negative impact of cybercrime, thus placing firms in a better position to use their data and create value; however, protection has a direct cost that hampers the stock of knowledge the firm can accumulate and, therefore, reduces the value of output as well. The market equilibrium analysis reveals that cyber-protection shifts both the demand curve and the supply curve down (for both trading positions of the two groups of producers), leading to the formation of a steady state with a lower equilibrium price.

Comparing the percentage changes in Table 4 with the changes in Table 3 (in this case, the variation from the no crime to the crime scenarios) one observes that, in what concerns stocks of knowledge and generated output, the decreases in each of the values is lower with protection than without it. Particularly relevant is the value of aggregate output, which, in the current case, is an intermediate value when compared with the other two. Specifically, for the two trading settings,

*a)* High data-intensity producers are data sellers and low data-intensity producers are data buyers:

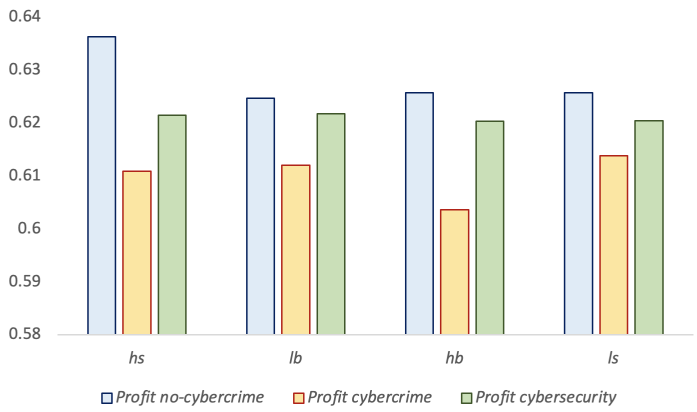
$$\left(Y^*_{cybercrime} = 0.6217\right) < \left(Y^*_{protection} = 0.6316\right) < \left(Y^*_{no-crime} = 0.6377\right)$$

*b)* High data-intensity producers are data buyers and low data-intensity producers are data sellers:

$$\left(Y^*_{cybercrime} = 0.6212\right) < \left(Y^*_{protection} = 0.6301\right) < \left(Y^*_{no-crime} = 0.6354\right)$$

Another way to compare and contrast the three scenarios, without cybercrime, with cybercrime, and with cyber-security, is to compare the steady-state profits of the two types of firms. Fig. 8 shows profits for high data-intensity firms and low data-intensity firms. Profits are highest when there is no cybercrime and when the high-data intensity firms sell data to low-data intensity buyers. Profits are lowest in the scenario of cybercrime with no cyber-protection. With cyber-protection, profits are at an intermediate level between no crime and crime with no protection. Protection mitigates the negative impact of cybercrime, thus placing firms in a better position to use their data and create value; however, protection has a direct cost that hampers profits.

Figure 8: Firm profits without/with cybercrime and cybersecurity




Legend: The figure shows profits for high data-intensity firms (which can be sellers, $hs$, or buyers, $hb$) and low data-intensity firms (which can be sellers, $ls$, or buyers, $lb$). Profits are highest when there is no cybercrime (in light blue) and when the high-data intensity firms sell data to low-data intensity buyers. Profits are lowest in the scenario of cybercrime with no cyber-protection (light yellow). With cyber-protection (light green), profits are at an intermediate level between no crime and crime with no protection.

## 2.3 Cybercrime driven innovation

Thus far, data has been interpreted as a prediction device. Prediction is subject to diminishing returns in the sense that improving it makes the production techniques to progressively approach a fixed maximum efficiency level. To generate endogenous growth in a model with data, the literature suggests various pathways. In Jones and

Tonetti (2020) data is an input employed to improve the quality of ideas, i.e., data is a driver of increasing quality. In Cong et al. (2021), Cong et al. (2022), Xie and Zhang (2022), and Hou et al. (2022), data is, instead, a driver of expanding varieties; data is employed by a research sector to expand the innovation possibilities frontier under the form of a larger number of varieties. The economy can also grow due to the accumulation of data, with data used as any other input directly in production (Freeman et al. (2021)).

To introduce dynamics and growth in the explored setting, we choose a different strategy, linking growth directly to the cyber-security measures that firms take when threatened by cybercrime. Cyber-security may prevent cyber-attacks to continue to take place with the same severity in the specific firm under attack, but it can also lead to innovation, making new production lines to emerge and, therefore, leading to the creation of new firms. To explore this process, consider that cybercrime spreads over firms through a simple diffusion mechanism. Let this mechanism be modelled in the following way,

$$I_{i,\vartheta,t+1} - I_{i,\vartheta,t} = \alpha_i \left(1 - \frac{I_{i,\vartheta,t}}{I_{i,t}}\right) I_{i,\vartheta,t}, \quad I_{i,\vartheta,0} > 0 \text{ given}, \alpha > 0, i = l, h \tag{26}$$

In equation (26), $I_{i,t}$ is the total number of producers at date $t$ that belong to group $i$, and $I_{i,\vartheta,t}$ is the number of producers subject to cyber-attacks at the same date, for firms in the same group. At some initial steady state date, $I_{l,0} = 1 - u$, $I_{h,0} = u$, and $I_0 = I_{l,0} + I_{h,0} = 1$. Equation (26) considers two motives for cyber-attacks: opportunity and attractiveness. Opportunity is translated in the convergence process implicit in the equation: the lower the number of firms that have already been attacked, the faster is the convergence of $I_{i,\vartheta,t}$ towards $I_{i,t}$. Attractiveness is linked to the gain accomplished by criminals when performing attacks. Recall that this gain corresponds to the data that criminals are able to subtract from firms, in an amount $\vartheta\left(z_i + \Delta_{i,t}\delta_{i,t}\right)$. Therefore, we postulate that the speed of diffusion of cybercrime depends on the value of this amount of data, such that:

$$\alpha_i = \alpha_0 \vartheta\left(z_i + \Delta_i^* \delta_i^*\right), \quad \alpha_0 > 0 \tag{27}$$

Take into consideration that equation (26) is applicable to four different cyber-attack processes: it applies to firms in each of the two assumed business sectors (low

data-intensity and high data-intensity) and to the two trading scenarios (in which firms in the two sectors are, alternatively, sellers or buyers of data).

According to the reasoning in the previous section, investment in cyber-security might be undertaken to guarantee partial protection against cyber attacks. The cost of this investment is proportional to the amount of data requiring protection. The question one might raise at this point is when will firms react to cyber attacks. We consider that the reaction to cybercrime is heterogeneous across firms, with some firms reacting immediately and others adopting a sluggish response. The following equation captures this idea,

$$I_{i,\tau,t+1} = \sum_{j=0}^{\infty} \lambda_i (1-\lambda_i)^j I_{i,\vartheta,t-j}, \quad \lambda_i \in (0,1), i = l, h \tag{28}$$

In equation (28), variable $I_{i,\tau,t}$ represents the universe of protected firms in sector $i$. The number of protected firms at $t+1$ is a weighted average of firms that have been attacked in all prior moments and that decide to adopt protection at the mentioned period. Variable $\lambda_i$ measures the degree of inertia in adopting protection; if the value of $\lambda_i$ is close to 1, the large majority of producers subject to cyber attacks at time $t$ will adopt protection at $t+1$; if the value of $\lambda_i$ is close to zero, the promptness in reacting to data breaches is more contained. One should expect firms with stronger losses on their stocks of knowledge, when attacked, to be more concerned in protecting data, and therefore, we consider that $\lambda_i$ depends on the difference $\Omega^*_{i,\tau} - \Omega^*_{i,\vartheta}$. Specifically, the following functional form is taken:

$$\lambda_i = \frac{\Omega^*_{i,\tau} - \Omega^*_{i,\vartheta}}{\lambda_0 + (\Omega^*_{i,\tau} - \Omega^*_{i,\vartheta})}, \quad \lambda_0 > 0 \tag{29}$$

Parameter $\lambda_0$ measures the sensitivity of the difference between stocks of knowledge (with and without protection) over the degree of protection inertia: the lower the value of this parameter, the smaller is the degree of sluggishness (i.e., the closer $\lambda_i$ will be to 1).

To complete the diffusion framework, one needs to address the evolution of the number of firms. The number of firms grows with innovation, i.e., with the number

of produced varieties. The adopted assumption is that cyber-security may have a side effect: as firms invest in cyber-security they may discover, with some probability $p_i$, new prototypes, that will lead to the creation of new production units, which will produce the new varieties. This idea is translated into the following equation:

$$I_{i,t+1} = I_{i,t} + p_i I_{i,\tau,t}, \quad i = l, h \tag{30}$$

Given a large number of firms, at each period, when $I_{i,\tau,t}$ firms invest in cyber-security, there is a share of these firms, $p_i I_{i,\tau,t}$, that will innovate, creating new varieties that will be added to the stock of already existing businesses. Probability $p_i$ is contingent on the investment in cyber-security; the higher the value of this investment, the higher will be the probability of innovation. Specifically, let:

$$p_i = \phi \frac{\tau_i^*}{1+\tau_i^*}, \quad \tau_i^* = \widetilde{\tau}\left(z_i + \Delta_i^* \delta_i^*\right), \phi \geq 0 \tag{31}$$

In expression (31), the value of parameter $\phi$ must be such that $p_i \leq 1$.

Equations (26), (28), and (30) compose a three dimensional system with three endogenous variables, with all of them representing numbers of firms: the total number of firms, $I_{i,t}$, the number of infected firms, $I_{i,\vartheta,t}$, and the number of protected firms, $I_{i,\tau,t}$. From these three groups, it is straightforward to compute the amount of firms that have not yet been attacked, $I_{i,t} - I_{i,\vartheta,t}$, and the number of firms that have been attacked but are not yet protected, $I_{i,\vartheta,t} - I_{i,\tau,t}$. These two differences are important to calculate the value of aggregate output. Note that the evolution of the number of firms, in each of the mentioned categories, is determined by three parameters, whose values we have attached to the growth-data framework: the rate of diffusion, $\alpha_i$, the promptness in protection, $\lambda_i$, and the probability of innovation, $p_i$.

Once the shares of producers in each position have been derived, one can compute the value of output and the respective growth rate. Because new varieties are generated with a constant probability, one expects the growth rate of output to converge as well to a balanced growth path, where the growth rate is positive and constant. This will be confirmed through the numerical example that is presented in the end of the section. The level of output, for each of the two groups of firms comes,

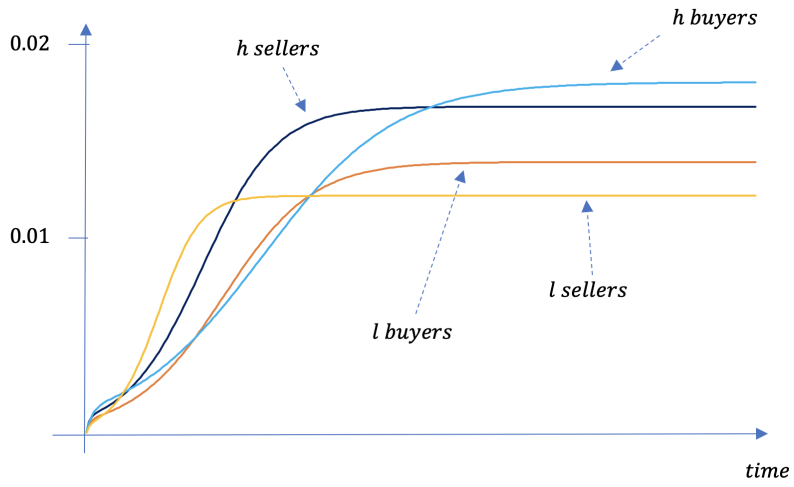
$$Y_{i,t}^{*} = (I_{i,t} - I_{i,\vartheta,t})A_{i}^{*} + (I_{i,\vartheta,t} - I_{i,\tau,t})A_{i,\vartheta}^{*} + I_{i,\tau,t}A_{i,\tau}^{*}, \quad i = l, h \tag{32}$$

Steady state values $A_{i}^{*}, A_{i,\vartheta}^{*}, A_{i,\tau}^{*}$ represent the level of output of each firm with no cybercrime, with cybercrime and with cyber-security, respectively. The total output in the economy amounts to:

$$Y_{t}^{*} = Y_{l,t}^{*} + Y_{h,t}^{*} \tag{33}$$

To better understand the proposed mechanism, consider a numerical example. Take the same parameter values as in previous sections and add the following: $\alpha_0 = 0.25$, $\lambda_0 = 0.5$, and $\phi = 2$. Fig. 9 displays the growth rate of the number of firms in each sector, for each one of the two trading possibilities. In the long-term, the number of firms grows at a constant rate, with the number of producers in the $h$ group growing at a faster rate than the number of producers in the $l$ group.

Figure 9: Trajectories of the growth rates of the number of firms

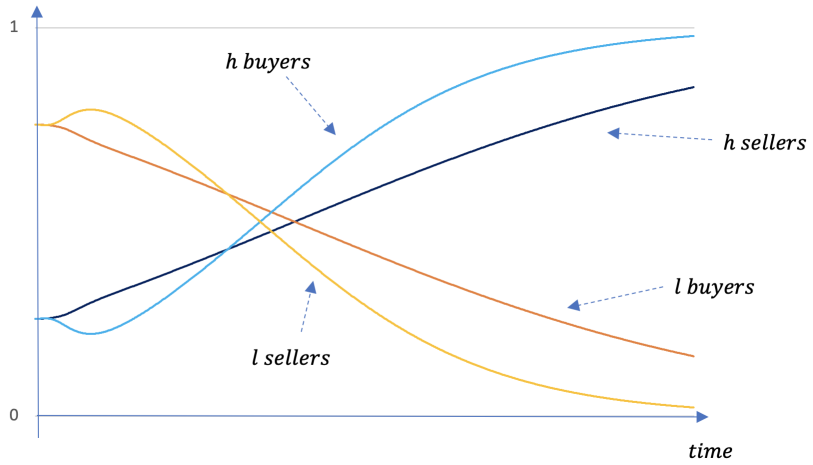



Legend: The figure displays the growth rate of the number of firms in each sector, for each one of the two trading possibilities. In the long-term, the number of firms grows at a constant rate, with the number of producers in the $h$ group growing at a faster rate than the number of producers in the $l$ group.

Although the number of firms in both groups increases over time, the number of high data-intensity firms grows faster, meaning that these producers will end up by

being largely dominant in the economy. This effect is evidenced in Fig. 10, which displays the evolution of the shares of firms in each sector. Starting at $u = 0.25$, this share will asymptotically converge to 1, whether high data-intensity firms are data sellers or data buyers.

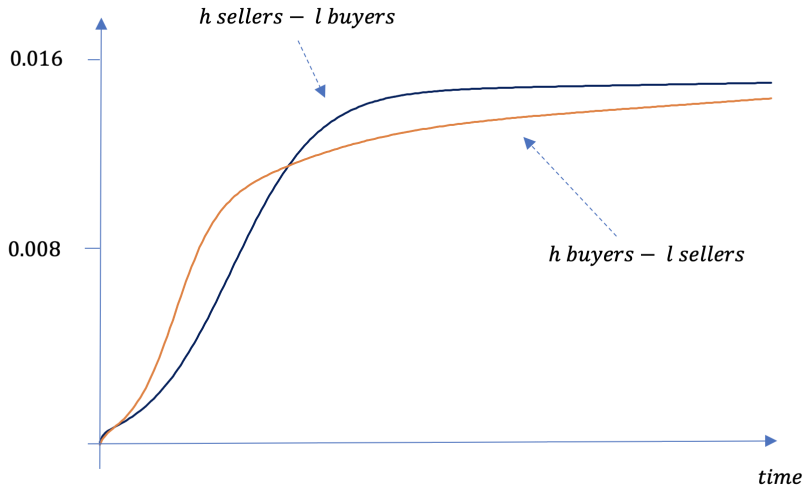
Figure 10: Trajectories of the shares of firms in each production sector




Legend: The figure displays the evolution of the shares of firms in each sector. Although the number of firms in both groups increases over time, the number of high data-intensity firms grows faster, meaning that these producers will end up by being largely dominant in the economy. Starting at $u = 0.25$, the share of high data-intensity firms will asymptotically converge to 1, independently on whether high data-intensity firms are data sellers or buyers.

We can now apply equations (32) and (33) to compute the level of output of the economy and the corresponding growth rate. Fig. 11 reveals that the growth rate will converge to a constant positive value, for both trading possibilities.

Figure 11: Trajectory of the growth rate of output




Legend: The figure displays the growth rate of output in the two production scenarios.

### 2.3.1 The Analytics of the BGP

Fig. 9 revealed, for the proposed numerical example, that the growth rate of the number of firms (i.e., the growth rate of the number of produced varieties) converges to a constant positive value: the suggested mechanism leads to the formation of a BGP where growth is sustained over time. The source of sustained growth is the innovation that emerges when producers invest in cyber-security. The mentioned rate can be analytically derived, as well as the growth rate of output.

Designate the long-term constant growth rate of $I_{i,t}$ by $\gamma_i$. Difference equation (30) reveals that if $I_{i,t}$ grows at constant rate $\gamma_i$, then ratio $I_{i,\tau,t}/I_{i,t}$ must be constant, and therefore the BGP growth rate of $I_{i,\tau,t}$ is also $\gamma_i$. Furthermore, given the cybercrime diffusion process, (26), one realizes that after the transient phase is overcome, also $I_{i,\vartheta,t}$ will grow at rate $\gamma_i$. Hence, in this scenario, the ratio between firms under attack and total number of firms is, given (26),

$$\frac{I_{i,\vartheta,t}}{I_{i,t}} = 1 - \frac{\alpha_i}{\gamma_i} \tag{34}$$

The BGP ratio between number of protected firms and the total number of firms is straightforward to draw from expression (30),

$$\frac{I_{i,\tau,t}}{I_{i,t}} = \frac{\gamma_i}{p_i} \tag{35}$$

Finally, note that under the constant growth rate assumption, equation (28) is equivalent to:

$$(1+\gamma_i)I_{i,\tau,t} = \sum_{j=0}^{\infty} \lambda_i \left(\frac{1-\lambda_i}{1+\gamma_i}\right)^j I_{i,\vartheta,t} \Rightarrow \frac{I_{i,\tau,t}}{I_{i,\vartheta,t}} = \frac{\lambda_i}{1+\gamma_i} \sum_{j=0}^{\infty} \left(\frac{1-\lambda_i}{1+\gamma_i}\right)^j = \frac{\lambda_i}{\lambda_i+\gamma_i} \tag{36}$$

Putting together expressions (34), (35), and (36), one obtains a system of equations that can be solved with respect to the growth rate $\gamma_i$. Two solutions are obtained, but only one corresponds to a positive quantity. This solution is:

$$\gamma_i = \frac{\sqrt{\left[\lambda_i\left(\alpha_i+p_i\right)\right]^2 + 4\alpha_i^2 p_i \lambda_i} - \lambda_i\left(\alpha_i+p_i\right)}{2\alpha_i} \tag{37}$$

The rate $\gamma_i$ in (37) can, in fact, be four different values, for the number of firms in the high data-intensity sector and for the number of firms in the low data-intensity sector, in each of the two possible trading positions. These values are the ones to which the trajectories in Fig. 9 converge to. Observe that the long-term growth rate of the number of production units is exclusively determined by the three entities that shape the cybercrime / cyber-security diffusion process: $\alpha_i$, $\lambda_i$, and $p_i$. Recall, from expressions (27), (29), and (31), that these values all depend on the equilibrium results of the initially explored data-growth model. Therefore, the growth rate of the number of varieties (which, as remarked below, is also the growth rate of output) is, basically, driven by data endowments, data trading, and the stocks of knowledge firms are able to extract and accumulate from data. Hence, the driver of sustained growth is, in fact, in this setting, the access to big data and its use as an input in production; although growth is triggered by the innovation originating in the investment in cyber-security, the essence of the growth process comes from the fact that all the elements $\alpha_i$, $\lambda_i$, and $p_i$ depend on the accumulation of data.

From equation (32), it is straightforward to conclude that if the output of a firm in each scenario (no crime, crime, security) is constant, and that if the number of firms

in each scenario (no crime, crime, security) grows at the same constant rate, then the aggregate output of firms in sector $i = l, h$ must also be $\gamma_i$. The growth rate of the aggregate output of the economy is a weighted average of the growth rate in each sector, i.e., given (33),

$$\frac{Y^*_{t+1} - Y^*_t}{Y^*_t} = \frac{\gamma_l Y^*_{l,t} + \gamma_h Y^*_{h,t}}{Y^*_{l,t} + Y^*_{h,t}} \tag{38}$$

which is also a constant value.

# 3 Empirical Analysis

A key result of our model is that firms can effectively address the negative consequences of cyber risk through innovation. In this section, we empirically examine this proposition. Specifically, we investigate whether firms facing higher cyber risk demonstrate greater innovation. To quantify the mechanisms discussed in our paper, we analyze whether firms with elevated cyber risk engage in innovative practices related to cyber security. Additionally, we explore whether these firms exhibit higher overall innovation levels, encompassing both cyber security and non-cyber-security domains. We further test whether the data intensive firms are the main drivers of our results.

Previous studies have primarily focused on developing cyber risk measures and examining their correlation with stock market returns (Florackis et al., 2023; Jamilov et al., 2021; Jiang et al., 2020). Other researchers have examined the characteristics of hacked firms, as well as the impact on their financial performance and risk management practices (Ettredge et al., 2018; Kamiya et al., 2021). Our specific objective is to comprehend the relationship between cyber risk and a firm's innovation.

## 3.1 Data

Answering our question requires two things. First, a measure of firm level cyber risk. Second, a measure of innovation at the firm level.

We source our measure of cyber risk for US-based publicly listed firms from Florackis et al. (2023). The authors have designed a cyber security risk score based on a textual analysis of the annual 10-K filings of these companies. For any given year, a firm's risk measure is derived from the similarity between the language used to detail *risk-factors* in its current-year 10-K filings and the *previous-year* 10-K filings of

a chosen 'training' set of firms. The firms in this training set are those that endured actual cyber attacks in the same year. The assumption is that firms that have fallen prey to actual cyber attacks likely had existing vulnerabilities, which would have been reflected in their risk disclosures in the previous year. As such, if a firm's language in its risk-factor disclosures strongly resembles the previous-year risk disclosures of firms that were indeed attacked, it is inferred to bear a high cyber security risk. The similarity score, which also serves as the cyber risk score, ranges from zero to one, with a higher score indicating a greater cyber risk. These cyber risk scores are available for the period from 2007 to 2018. Accordingly, we calculate all other remaining variables within this same period.

We capture innovation in various complementary ways. The first measure we use is the knowledge capital accumulation calculated by Ewens et al. (2020). Knowledge capital is the stock of research and development (R&D) expenditure net of the knowledge capital depreciation. Knowledge asset can also be thought of as an input to innovation, rather than output, as it represents expenditure on producing innovation. Our next set of measures explicitly capture innovation output.

Firms' patent activity represent their innovation output. Following the literature on innovation, we count patents filed by the firms by taking into account their scientific and economic value (Kogan et al., 2017; Aghion et al., 2013; Howell, 2017). In our first patent measure, we count number of patents filed by the weighing it with the number of forward citations it receives. The idea is that the more important a patent is scientifically, the more citations it receives (Hall et al., 2005; Kogan et al., 2017). Following the best practice in the literature, we adjust the count for the truncation bias. As the citations occur over time, a simple counting of cites will underestimate the importance of the patents that were issued towards the end of our sample period (Lerner and Seru, 2022; Dass et al., 2017; Hall et al., 2001). We correct for that using the well-established methodology proposed by Hall et al. (2001).

We also calculate value-weighted count of number of patents filed. We do so by weighing each patent by the economic value it creates. The economic value of a patent is the dollar amount of wealth generated for the patenting firm's shareholders, calculated from the stock market response to the news about the patent award. We scale the patent value by the firm's total assets, following Kogan et al. (2017).

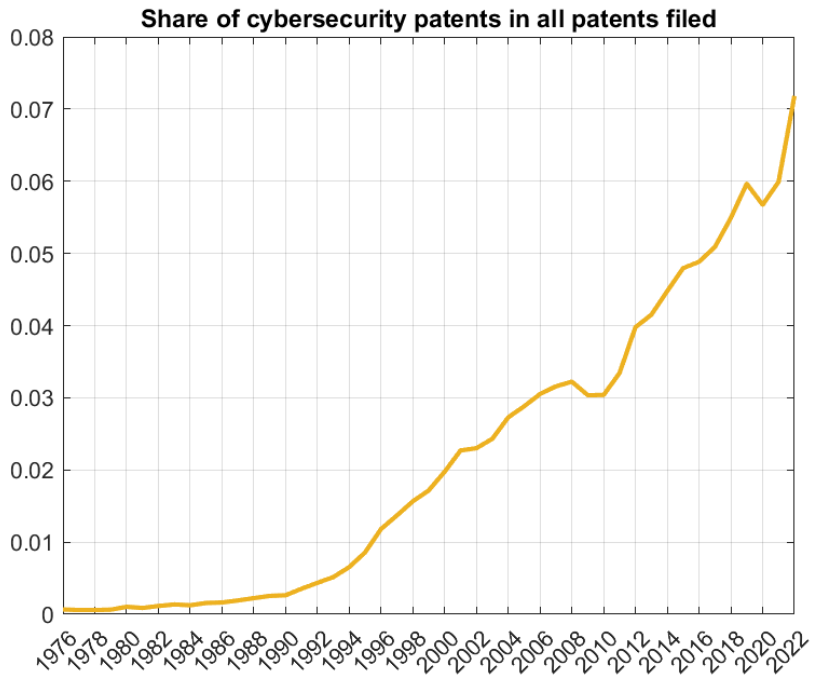
In an additional analysis, we examine whether firms exposed to more risks expand their areas of innovation. To do that, we extract the Cooperative Patent Classification (CPC) code for each filed patent. We then count the number of unique 'fields' in which

a firm files patents in a year. We define number of fields at different level of coarseness. A CPC code consists of five hierarchical parts: section, class, sub-class, group, and sub-group. Section is the highest level in hierarchy, and the most aggregative level, followed by class, subclass, and so on. For our purpose, we define patent fields at three alternative levels: section, class, and sub-class. We do not differentiate patents along the group or subgroup levels because we want to make sure that we are counting patent fields that are somewhat distinct from each other.

All our patent data is from the publicly available database maintained by the authors of Kogan et al. (2017).

We gauge cyber security innovation using the citation-weighted and value-weighted count of cyber security patents a firm files within a year. A patent is classified as a cyber security patent if the USPTO assigns it CPC codes associated with cyber security. For instance, CPC code G06F21/ is titled "Security arrangements for protecting computers, components thereof, programs or data against unauthorised activity". Our cyber security patent measure indicates a consistent growth in cyber security innovation over time, currently accounting for approximately 7

Figure 12: Cyber security innovation




To identify data-intensive firms, we create a measure grounded on two fundamental

premises. First, we propose that firms involved in the creation of AI technology are, by nature, data-intensive. Second, we posit that any firm, including those not directly engaged in AI development, can be considered data-intensive if the language used to describe its business mirrors that of AI-creating firms. In accordance with these premises, our measure is crafted in two steps. For the initial step, we utilize a newly published dataset by the USPTO, a product of their internal research, which classifies AI patents within the entire spectrum of patents filed at the USPTO Giczy et al. (2022). This helps us identify those US public firms that have submitted AI patent applications. In the second step, we employ a dataset curated by Hoberg and Phillips, which quantifies the textual similarity in the Business Description between any two firms (Hoberg and Phillips, 2016). The underlying principle here is that data-intensive firms are likely to portray their businesses in a similar light. Therefore, a firm not holding an AI patent is also considered data-intensive if its business description more closely aligns with those firms possessing AI patents.

We obtain firm level financial information from the merged CRSP-Compustat database. We calculate various financial variables and ratios to use them as control variables in our baseline regressions. Specifically, we use the following variables as controls: log of total assets, tobin's Q, asset tangibility, book-to-market ratio, cash-to-asset ratio, leverage, and return on assets. We winsorize all the variables at 0.5% on both sides of the distribution.

Table 1 presents summary statistics on our cyber risk and innovation measures. We see that more than a quarter of the firms do not face cyber risk. Further, as is well-known innovation activity is quite skewed. For instance, more than 50 percent of firm-years do not record any positive knowledge capital accumulation or any patent activity.

## 3.2 Empirical strategy

We conduct regression analysis to uncover the relationship between cyber risk and innovation. We rely on two aspects of our regression specification to identify the causal relation between cyber risk and innovation. First, we regress innovation measures on the lagged value of cyber risk score. Doing so addresses the simultaneity concerns. Second, we include firm fixed effects to absorb time invariant characteristics of firms that might affect this relationship. Moreover, we include year fixed effects to absorb shocks occurring over time and that are common across firms. Finally, we control for

Table 1: Summary statistics on cyber-risk score and innovation variables

| | N | mean | sd | p10 | p25 | p50 | p75 | p90 | p99 |
|---|---|---|---|---|---|---|---|---|---|
| Cyber-risk score | 44972 | 0.2 | 0.2 | 0 | 0 | 0.3 | 0.4 | 0.5 | 0.6 |
| Log(Knowledge stock) | 41479 | 1.6 | 2.2 | 0 | 0 | 0 | 3.4 | 5.0 | 8.0 |
| Log(R&D expenditure) | 44972 | 1.3 | 1.9 | 0 | 0 | 0 | 2.6 | 4.2 | 7.2 |
| Patents filed: simple count | 44972 | 9.2 | 49.9 | 0 | 0 | 0 | 0 | 8.0 | 291.0 |
| Patents filed: citation-weighted count | 44972 | 18.3 | 100.4 | 0 | 0 | 0 | 0 | 15.3 | 549.4 |
| Patents filed: value-weighted count | 44881 | 0.05 | 0.20 | 0 | 0 | 0 | 0 | 0.11 | 1.17 |
| Number of patent sections | 10616 | 3.3 | 2.0 | 1 | 2 | 3 | 4 | 6 | 9 |
| Number of patent classes | 10616 | 7.5 | 10.3 | 1 | 2 | 4 | 8 | 17 | 57 |
| Number of patent subclasses | 10616 | 14.1 | 25.0 | 1 | 3 | 6 | 14 | 31 | 145 |

N refers to the total number of firm-year. p10-p99 refer to the 10th to 99th percentile values. Cyber risk score lies between zero and one, with higher values indicating higher risk. Cyber risk score measure is obtained from Florackis et al. (2023). Knowledge stock is based on the estimates of knowledge stock net of knowledge depreciation from Ewens et al. (2020). Simple patent count refers to number of patents filed by the firm in a year. Citation-weighted patent count weighs each patent with the forward citation the patent receives, adjusting for the filing vintage. Value-weighted patent count is the sum of stock market value generated over all the patents filed by a firm in a year, scaled by total assets. Number of patent sections refers to the number of unique CPC sections associated with all the patent the firm files in a year. Similar explanation applies to patent classes, and subclasses, respectively.

various financial factors.

As visible from Table 1, our innovation variables have a right skew and contain high share of zeros. Therefore, applying ordinary least squares (OLS) estimation in a regression of the patent counts might result in inefficient parameter estimates. One possible solution could be using OLS estimation after a log transformation of our patent count variables. However, given a large number of zeros, a log transformation excludes substantial number of observations when estimating log-linear regressions. More importantly, log-linear regressions may even produce inconsistent estimates of the parameters (Silva and Tenreyro, 2011). Alternatively, we could log transform after adding one to each patent count, or apply inverse hyperbolic sine transformation. These transformations would retain zeros, however, they may also produce inconsistent estimates. Moreover, they may even have the opposite sign of the true relationship, as shown by Cohn et al. (2022).[5]

Econometricians recommend Poisson model to explicitly take into account many zeros and the right skew of the dependent variables. Because, in such a setting too, a Poisson model produces consistent estimators without requiring any assumptions about higher order model error moments (Cohn et al., 2022). In addition, and importantly

[5]Though, less fatal than other flaws the parameter estimates are also hard to interpret after the transformations (Cohn et al., 2022; Silva and Tenreyro, 2006).

for us, Poisson regression allows for separable group fixed effects (Correia et al., 2020; Cohn et al., 2022). Moreover, even though the Poisson model is generally considered to be useful for count data (such as patents), actually, it is valid even when the dependent variable is continuous with a non-negative domain (such as knowledge asset) (Silva and Tenreyro, 2011; Wooldridge, 1999).[6]

To study the relationship between the lagged value of cyber risk score ($\text{crscore}_{it-1}$) and innovation measure ($\text{innovation}_{it}$) we fit the following conditional expectation of an innovation measure that follows a Poisson distribution:

$$\mathbb{E}[\text{innovation}_{it}|\text{crscore}_{it-1}, \mathbf{x}_{it-1}, \eta_i, \tau_t] = \exp\left(\beta_c \text{crscore}_{it-1} + \beta \mathbf{x}_{it-1} + \eta_i + \tau_t\right) \quad (39)$$

where $\text{crscore}_{it-1}$ is the lagged value of cyber-risk score, $\mathbf{x}_{it-1}$ are *lagged* control variables, including size (log of total assets), Tobin's Q, asset tangibility, book-to-market ratio, cash-to-asset ratio, leverage, and return on assets. $\eta_i$ is the firm fixed effect, and $\tau_t$ is the year fixed effect.

We perform Poisson pseudo-maximum likelihood estimation to estimate the parameters of the model in (39).

We also study whether cyber risk score affects the *R&D productivity*. To do that, we follow Aghion et al. (2013), and in some specifications of (39) include *R&D stock* as a right hand side variable. In such specifications, the coefficient $\beta_c$ tells us whether firms with higher score innovate more per dollar of R&D stock. In specifications, where *R&D stock* is not included as a control variable, $\beta_c$ contains the effect of R&D productivity and additional effect of higher cyber risk on innovation.

Finally, we cluster standard errors at the firm level, to take into account the possibility of autocorrelation and hetereskedasticity in the error terms. Clustered standard errors are additionally useful because they are also robust to 'overdispersion' (and 'underdispersion') issues countered in Poisson regression (Cohn et al., 2022; Wooldridge, 1999).

## 3.3 Baseline results

In what follows, we report the results from our preferred Poisson estimation. We also report estimates from OLS regression of our innovation measures.[7]

[6] Well-known works employing Poisson regression with patent data include Azoulay et al. (2019); Aghion et al. (2013); Amore et al. (2013); Blundell et al. (1999); Hausman et al. (1984).

[7] Although, fully recognizing that this might not be the correct model specification.

Table 2 presents the results of regressing knowledge capital and R&D stock on lagged cyber-risk score. We find that firms accumulate more knowledge capital and R&D stock in response to a rise in cyber risk. Although, in the regression of knowledge capital, the Poisson model does not give a significant coefficient for cyber risk at the conventional 10% significance level, it is quite close. Moreover, the results are also confirmed by the regression of R&D stock, which shows a significant rise. The increase is also economically meaningful. For instance, one standard deviation change in cyber risk would lead to an increase in R&D by about 3% $[= 0.22(e^{0.124} - 1)]$, keeping everything else the same.

How do firms respond with their patenting output when they face a higher cyber risk? Do they file more patents because they accumulate more R&D stock, or do they also respond by increasing their R&D productivity? To test that, we regress patent-count variables on lagged cyber-risk in Table 3 and Table 4. The first two columns of both the tables exclude R&D as a control variable. Therefore, these specifications test the change in firm's innovation output to cyber risk. The change includes the effect of cyber risk on innovation input, as well is its effect on the R&D productivity. In columns (3) and (4) we also include the stock of R&D capital as an explanatory variable. Therefore, the coefficient on cyber-risk score give us the estimate of how in response to an increase in cyber risk, a firm's patent count changes keeping its innovation input (R&D capital) unchanged.

Table 3 presents the regression results with citation-weighted patent count as the dependent variable. The first observation we make is that in all the specification in the table, the coefficient on Cyber-risk score is positive, indicating that firms patent more in response to a cyber-risk shock. From our Poisson estimate in column (2), we can quantify the effect. A one standard deviation increase in cyber risk in a year leads the firm to file 5% $[= 0.22(e^{0.201} - 1)]$ more patents the next year. We observe from column (4) that firms file more patents per dollar of R&D stock. We estimate that in response to a one standard-deviation shock in cyber risk, firm's R&D productivity rises by about 4.2%.

We arrive at similar conclusion when we use value-weighted patent count in Table 4. A one standard-deviation shock in cyber risk leads to the firm filing about 7% more patents in value-weighted terms (column 2). Out of this increase, about 6% is due to the increase in R&D productivity (column 4).

Table 2: Regression of knowledge stock and R&D stock

| | Knowledge stock | | R&D stock | |
|---|---|---|---|---|
| | OLS | Poisson | OLS | Poisson |
| | (1) | (2) | (3) | (4) |
| Cyber-risk score | 21.38** | 0.0868 | 9.755** | 0.124* |
| | (9.223) | (0.0542) | (4.468) | (0.0659) |
| ln(Asset) | 60.23*** | 0.486*** | 32.03*** | 0.551*** |
| | (9.959) | (0.0381) | (5.202) | (0.0339) |
| Tobin's Q | 2.894** | 0.0300*** | 2.557*** | 0.0537*** |
| | (1.359) | (0.00622) | (0.697) | (0.00773) |
| Tangibility | -2.051 | 0.484* | -9.190 | 0.286 |
| | (30.26) | (0.253) | (13.67) | (0.271) |
| Book-to-market | -0.781 | -0.0224** | 0.319 | 0.00190 |
| | (1.013) | (0.0109) | (1.018) | (0.0413) |
| Cash-to-asset | -45.29*** | -0.169* | -24.91*** | -0.195* |
| | (16.80) | (0.0895) | (6.908) | (0.106) |
| Leverage | -5.477 | -0.150* | -3.998 | -0.212** |
| | (15.01) | (0.0798) | (6.407) | (0.0841) |
| ROA | -45.97*** | -0.216*** | -18.87*** | 0.00708 |
| | (11.22) | (0.0716) | (5.475) | (0.0749) |
| Firm FE | Yes | Yes | Yes | Yes |
| Year FE | Yes | Yes | Yes | Yes |
| N | 31601 | 14921 | 34592 | 15038 |

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. Standard errors are in parentheses. Standard errors are clustered at the firm level. N refers to the total number of firm-year. Cyber score, and control variables are lagged by one year. Cyber-risk score measure is obtained from Florackis et al. (2023). Knowledge stock is based on the estimates of knowledge stock net of knowledge depreciation from Ewens et al. (2020). Other control variables are computed using WRDS CRSP-Compustat merged data. Tobin's Q is defined as Total assets (at) minus common equity (ceq) plus market value of equity (prcc_f × csho), as a ratio of total assets (at). ROA is defined as operating income before depreciation (oibdp) to total assets (at). Tangibility is defined as total property, plant and equipment (ppent) scaled by total assets (at). Leverage is long-term debt (dltt) plus debt in current liabilities (dlc), as a ratio of total assets (at). Book-to-market ratio is book value of common equity (ceq) divided by the market value of common equity (prcc_f × csho). Cash-to-asset is the ratio of cash and short-term investments (che) to total assets (at).

Table 3: Regression of citation-weighted patent count

| | Citation-weighted patent count | | | |
|---|---|---|---|---|
| | OLS | Poisson | OLS | Poisson |
| | (1) | (2) | (3) | (4) |
| Cyber-risk score | 4.060* | 0.201** | 3.784* | 0.176* |
| | (2.135) | (0.101) | (2.138) | (0.0994) |
| ln(Asset) | 2.064** | 0.141*** | 1.190* | 0.0352 |
| | (0.824) | (0.0493) | (0.716) | (0.0516) |
| Tobin's Q | 0.282 | 0.0159 | 0.268 | 0.0139 |
| | (0.283) | (0.0153) | (0.282) | (0.0154) |
| Tangibility | 3.032 | 0.0758 | 2.834 | -0.0657 |
| | (5.990) | (0.559) | (5.979) | (0.538) |
| Book-to-market | -0.00283 | 0.0186 | 0.0392 | 0.0233 |
| | (0.247) | (0.0516) | (0.248) | (0.0516) |
| Cash-to-asset | -3.186 | -0.00373 | -2.717 | 0.0565 |
| | (2.590) | (0.150) | (2.570) | (0.148) |
| Leverage | -4.420* | -0.163 | -4.257* | -0.107 |
| | (2.297) | (0.189) | (2.295) | (0.191) |
| ROA | -0.371 | 0.105 | 0.603 | 0.151 |
| | (1.295) | (0.184) | (1.289) | (0.181) |
| ln(R&D stock) | | | 2.978*** | 0.190*** |
| | | | (0.757) | (0.0441) |
| Firm FE | Yes | Yes | Yes | Yes |
| Year FE | Yes | Yes | Yes | Yes |
| N | 34592 | 12900 | 34592 | 12900 |

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. Standard errors are in parentheses. Standard errors are clustered at the firm level. Cyber score, and control variables are lagged by one year. Cyber-risk score measure is obtained from Florackis et al. (2023). Citation-weighted patent count weighs each patent with the forward citation the patent receives, adjusting for the filing vintage. For the description of control variables, see notes for Table 2.

Table 4: Regression of value-weighted patent count

| | Value-weighted patent count | | | |
|---|---|---|---|---|
| | OLS | Poisson | OLS | Poisson |
| | (1) | (2) | (3) | (4) |
| Cyber-risk score | 0.0182*** | 0.277** | 0.0186*** | 0.238** |
| | (0.00703) | (0.122) | (0.00699) | (0.116) |
| ln(Asset) | -0.0266*** | -0.301*** | -0.0252*** | -0.413*** |
| | (0.00449) | (0.0489) | (0.00406) | (0.0568) |
| Tobin's Q | 0.00197 | 0.00296 | 0.00199 | -0.00128 |
| | (0.00184) | (0.00836) | (0.00183) | (0.00826) |
| Tangibility | 0.0171 | 0.351 | 0.0173 | 0.135 |
| | (0.0211) | (0.455) | (0.0211) | (0.456) |
| Book-to-market | 0.00190** | -0.0961 | 0.00184* | -0.0955 |
| | (0.000943) | (0.0644) | (0.000941) | (0.0639) |
| Cash-to-asset | 0.0466*** | 0.468*** | 0.0459*** | 0.501*** |
| | (0.0161) | (0.160) | (0.0163) | (0.155) |
| Leverage | 0.0162 | -0.00786 | 0.0160 | 0.0386 |
| | (0.0116) | (0.111) | (0.0115) | (0.107) |
| ROA | 0.00119 | 0.0900 | -0.000269 | 0.202** |
| | (0.0113) | (0.0878) | (0.0111) | (0.0879) |
| ln(R&D stock) | | | -0.00446 | 0.178*** |
| | | | (0.00449) | (0.0530) |
| Firm FE | Yes | Yes | Yes | Yes |
| Year FE | Yes | Yes | Yes | Yes |
| N | 34579 | 12896 | 34579 | 12896 |

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. Standard errors are in parentheses. Standard errors are clustered at the firm level. N refers to the total number of firm-year. Cyber-risk score, and control variables are lagged by one year. Cyber risk score measure is obtained from Florackis et al. (2023). Value-weighted patent count is the sum of stock market value generated over all the patents filed by a firm in a year, scaled by total assets. For the description of control variables, see notes for Table 2.

## 3.4 Cyber risk and cyber security innovation

To thoroughly investigate the mechanisms underpinning the relationship between cyber security and innovation, we delve into the remaining parts of the loop we developed in theoretical framework earlier. Our initial inquiry centers around whether firms exposed to heightened cyber risk are more likely to increase their focus on cyber security innovation. Subsequently, we explore if an uptick in cyber security innovation could stimulate a broader surge in overall innovation. In the ensuing regression tables, we will restrict our focus to the pertinent coefficient estimates, suppressing those associated with the controls. Further, we now work exclusively with our preferred Poisson model.

Table 5 displays the regression of the number of filed cyber security patents against the lagged cyber risk, across different specifications. All regression models incorporate the controls used in previous analyses, log of R&D stock, and year fixed effects. Industry fixed effects are included in columns (1) through (4), while firm fixed effects are applied in specifications (5) and (6). The industry fixed effects are used in the initial specifications due to the fact that a relatively small number of firms file cyber security patents. This leads to a reduced number of observations if we apply firm fixed effects, which complicates the task of obtaining precise estimates. Within the first four specifications, we alternate between the exclusion and inclusion of the lagged count of both cyber security and overall patents.

Our analysis indicates that an increase in cyber risk prompts firms to file a greater number of cyber security patents. This positive effect is still evident in the most restrictive specification featuring firm fixed effects. However, due to the limited number of observations, we cannot assert our conclusions with complete confidence.

In order to examine the next segment of the loop, we investigate whether firms with a higher degree of innovation in cyber security also exhibit a greater level of overall innovation. We undertake regression analyses where we regress citation-weighted and value-weighted patent counts against both the lagged counts of cyber security patent filings and the lagged cyber risk scores (as presented in Table 6).

Columns (1) and (2) in Table 6 mirror our previous baseline findings. In the subsequent models, we also investigate the effect of the lagged counts of cyber security patents. We ascertain that an increase in cyber security patents leads to an overall surge in innovation, as reflected in both measures of innovation. To refine our results, we also account for the contemporaneous counts of cyber security patents in specifications (4)

Table 5: Regression of cyber security innovation

| | Cit-wtd CS patent # | | Val-wtd CS patent # | | Cit-wtd CS patent # | Val-wtd CS patent # |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| L.Cyber risk | 1.188*** | 0.693*** | 1.792*** | 1.187*** | 0.338 | 0.0257 |
| | (0.351) | (0.242) | (0.402) | (0.404) | (0.285) | (0.271) |
| L.# cit-wtd CS patent | No | Yes | No | No | No | No |
| L.# val-wtd CS patent | No | No | No | Yes | No | No |
| L.# cit-wtd patent | No | Yes | No | No | No | No |
| L.# val-wtd patent | No | No | No | Yes | No | No |
| Size + other controls | Yes | Yes | Yes | Yes | Yes | Yes |
| NAICS-3 FE | Yes | Yes | Yes | Yes | No | No |
| Firm FE | No | No | No | No | Yes | Yes |
| Year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| N | 29283 | 29283 | 29273 | 29273 | 3502 | 3501 |

Table 6: Regression of counts of patent filed

| | # Cit-wtd patent | # Val-wtd patent | # Cit-wtd patent | | # Val-wtd patent | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| L.Cyber risk | 0.199** | 0.276** | 0.0764 | 0.0716 | 0.220** | 0.233** |
| | (0.101) | (0.122) | (0.0728) | (0.0736) | (0.0991) | (0.0984) |
| L.# cit-wtd CS patent | | | 0.00286** | 0.00135* | | |
| | | | (0.00114) | (0.000802) | | |
| L.# val-wtd CS patent | | | | | 2.487*** | 1.607*** |
| | | | | | (0.554) | (0.591) |
| # cit-wtd CS patent | No | No | No | Yes | No | No |
| # val-wtd CS patent | No | No | No | No | No | Yes |
| L.# cit-wtd patent | No | No | Yes | Yes | No | No |
| L.# val-wtd patent | No | No | No | No | Yes | Yes |
| Size + other controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| N | 12900 | 12896 | 12900 | 12900 | 12896 | 12896 |

and (6). Our results maintain their significance and positive orientation. This suggests that, even when the number of cyber security patents is held constant, there is an increase in the total number of patents when firms engage in cyber security innovation. This implies that firms also augment their portfolio of non-cyber-security patents in response to cyber security innovation.

## 3.5 Data intensive firms and their response to cyber security risk

Next we study how this dynamic differs between the data-intensive and non-data intensive firms. Our model posits a feedback loop for the data economy, i.e. an economy reliant on the data that is subject to the risk of being stolen. We therefore, expect this mechanism to apply on data-intensive firms and not on the non-data-intensive firms.

We construct an dummy variable that takes value 1 if the firm is identified as a data intensive firm by our method described earlier. We then run the regressions similar to as in the previous section, however now we interact the lagged cyber security score with the dummy on data intensity. The results are presented in Table 7.

We find that even though the data intensive firms account only for a minority of the observations (roughly 40%), our baseline results are driven by them. Indeed, the regressions show that cyber risk score has even sometimes negative effects on innovation in the non-data intensive firms, although, the results are never significant.

Table 7: Regression with data intensity

| | Cit-wtd patent # | Val-wtd patent # | Cit-wtd patent # | | Val-wtd patent # | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| L.Cyber risk*(data int =0) | -0.0872 | 0.186 | -0.0607 | -0.0463 | 0.215 | 0.260 |
| | (0.156) | (0.191) | (0.109) | (0.111) | (0.157) | (0.162) |
| L.Cyber risk*(data int = 1) | 0.289** | 0.293** | 0.121 | 0.110 | 0.221** | 0.228** |
| | (0.115) | (0.126) | (0.0802) | (0.0807) | (0.104) | (0.103) |
| L.# cit-wtd CS patent | | | 0.00281** | 0.00131 | | |
| | | | (0.00114) | (0.000804) | | |
| L.# val-wtd CS patent | | | | | 2.487*** | 1.611*** |
| | | | | | (0.554) | (0.590) |
| # cit-wtd CS patent | No | No | No | Yes | No | No |
| # val-wtd CS patent | No | No | No | No | No | Yes |
| L.# cit-wtd patent | No | No | Yes | Yes | No | No |
| L.# val-wtd patent | No | No | No | No | Yes | Yes |
| Size + other controls | Yes | Yes | Yes | Yes | Yes | Yes |
| Firm FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| N | 12900 | 12896 | 12900 | 12900 | 12896 | 12896 |

## 3.6 Cyber risk and patent fields

Do firms broaden the fields in which they innovate in response to the cyber risk? To answer that, we regress the number of patent fields in a firm's patent filings on cyber risk scores (Table 8). We use different definitions of patent fields representing various levels of aggregation in CPC codes. Patent sections are at the top level, with different sections representing very distinct areas. Patent classes have smaller distinction across them, and so on.

We find that while point estimates are positive, they are not significant for section count or class count. For subclasses, there is a positive and significant effect of cyber risk when we estimate an OLS model. However, we cannot estimate it precisely with Poisson regression. Quantitatively, the change in number of subclasses in response to

Table 8: Regression of patent-field count

| | Count patent sections | | Count patent classes | | Count patent sub-classes | |
|---|---|---|---|---|---|---|
| | OLS | Poisson | OLS | Poisson | OLS | Poisson |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Cyber-risk score | 0.0566 | 0.0106 | 0.583 | 0.00726 | 1.958** | 0.0193 |
| | (0.116) | (0.0338) | (0.424) | (0.0484) | (0.960) | (0.0521) |
| ln(Asset) | 0.154*** | 0.0508*** | 0.839*** | 0.122*** | 2.086*** | 0.159*** |
| | (0.0502) | (0.0168) | (0.265) | (0.0351) | (0.735) | (0.0468) |
| Tobin's Q | 0.0118 | 0.00437 | 0.0133 | 0.00454 | 0.0365 | 0.00730 |
| | (0.00875) | (0.00318) | (0.0299) | (0.00543) | (0.0732) | (0.00690) |
| Tangibility | -0.0134 | 0.00267 | 1.837 | 0.250 | 5.774 | 0.402 |
| | (0.345) | (0.107) | (1.622) | (0.215) | (4.662) | (0.315) |
| Book-to-market | 0.0162 | 0.00482 | 0.124 | 0.00926 | 0.378 | 0.0105 |
| | (0.0426) | (0.0149) | (0.165) | (0.0241) | (0.418) | (0.0283) |
| Cash-to-asset | 0.0686 | 0.0257 | 0.626 | 0.0911 | 1.026 | 0.0684 |
| | (0.136) | (0.0469) | (0.443) | (0.0709) | (1.086) | (0.0851) |
| Leverage | -0.183 | -0.0609 | -0.254 | -0.0581 | -0.535 | -0.0729 |
| | (0.126) | (0.0438) | (0.401) | (0.0720) | (0.886) | (0.0876) |
| ROA | -0.00596 | 0.0117 | -0.478 | -0.000475 | -1.462* | -0.00603 |
| | (0.0870) | (0.0324) | (0.315) | (0.0519) | (0.798) | (0.0652) |
| ln(R&D stock) | 0.0946** | 0.0282** | 0.195 | 0.0426* | 0.160 | 0.0402 |
| | (0.0373) | (0.0124) | (0.144) | (0.0230) | (0.364) | (0.0296) |
| Firm FE | Yes | Yes | Yes | Yes | Yes | Yes |
| Year FE | Yes | Yes | Yes | Yes | Yes | Yes |
| N | 8641 | 8641 | 8641 | 8641 | 8641 | 8641 |

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. Standard errors are in parentheses. Standard errors are clustered at the firm level. N refers to the total number of firm-year. Cyber-risk score, and control variables are lagged by one year. Cyber risk score measure is obtained from Florackis et al. (2023). Number of patent sections refers to the number of unique CPC sections associated with all the patent the firm files in a year. Similar explanation applies to patent classes, and subclasses, respectively. For the description of control variables, see notes for Table 2.

a rise in cyber risk is positive and non-negligible. A one standard deviation shock in cyber risk leads to a 0.4% increase in patent fields when we define fields in terms of the count of patent subclasses.

Overall, we find some evidence that firms expand the areas of innovation in response to cyberrisk, even though we would be reluctant to place a lot of confidence in this finding.

### 3.7 Cyber risk and financial variables

Does a rise in cyber risk affect a firm's profitability? Cyber risk can reduce a firm's profitability by diverting its resources towards cyber protection measures. It might even go up if the higher innovation in response to cyber risk creates new profitable opportunities. However, the two forces might counteract each other as well.

Table 9 presents results of a set of regressions on different financial variables. The first column regresses return on assets (ROA) on lagged cyber risk measure and other controls. We find no negative effect of cyber risk on profitability, indicating that innovation helps firms to hedge their profits against cyber risk.

In a similar regression given in columns (2)-(4), we find no significant effect of cyber-risk shock on a firm's Tobin's Q, Book-to-market ratio, and Leverage.

## 4 Conclusion

This paper explores the relationship between cybercrime and digital innovation and their combined impact on economic growth. We construct a growth model of the data economy where data, crucial for business optimization, is at risk of damage by cyber criminals. Our framework shows that cybercrime causes lower growth and innovation in firms, but sustained growth is still possible through innovation that compensates for cybercrime's loss of knowledge. The increased threat of cybercrime also drives innovation in security measures and systems, leading to advancements in technology and long-term growth. Our empirical analysis confirms that firms respond to cyberrisk with a rise in R&D and patenting activity and no change in profitability, as the risk is offset by innovation.

Table 9: Regression of financial variables

| | ROA (1) | Tobin's q (2) | Book-to-market (3) | Leverage (4) |
|---|---|---|---|---|
| Cyber-risk score | 0.00885 | 0.0709 | -0.0161 | -0.00787 |
| | (0.00851) | (0.0621) | (0.0372) | (0.00813) |
| ln(Asset) | 0.0200*** | -0.403*** | 0.232*** | 0.0366*** |
| | (0.00543) | (0.0351) | (0.0184) | (0.00403) |
| Tobin's Q | 0.0152*** | | -0.0360*** | -0.00217 |
| | (0.00251) | | (0.00401) | (0.00179) |
| Tangibility | -0.0905*** | -0.326* | 0.282** | 0.0792*** |
| | (0.0312) | (0.184) | (0.113) | (0.0278) |
| Book-to-market | -0.0233*** | -0.167*** | | -0.0180*** |
| | (0.00282) | (0.0177) | | (0.00287) |
| Cash-to-asset | -0.126*** | 0.493*** | -0.116** | -0.108*** |
| | (0.0208) | (0.150) | (0.0525) | (0.0162) |
| Leverage | -0.0107 | 0.284** | -0.618*** | |
| | (0.0192) | (0.124) | (0.0574) | |
| ln(R&D stock) | -0.0245*** | 0.0482 | -0.0270* | 0.000962 |
| | (0.00496) | (0.0311) | (0.0138) | (0.00445) |
| ROA | | 0.0187 | -0.0782* | -0.0819*** |
| | | (0.121) | (0.0417) | (0.0138) |
| Firm FE | Yes | Yes | Yes | Yes |
| Year FE | Yes | Yes | Yes | Yes |
| N | 34591 | 34564 | 34564 | 34577 |

$^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$. Standard errors are in parentheses. Standard errors are clustered at the firm level. N refers to the total number of firm-year. Cyber-risk score, and control variables are lagged by one year. Cyber risk score measure is obtained from Florackis et al. (2023). ROA stands for return on assets. All estimations are based on OLS regression. For the description of variables, see notes for Table 2.

# References


**Aghion, Philippe, John Van Reenen, and Luigi Zingales**, "Innovation and institutional ownership," *American Economic Review*, 2013, *103* (1), 277–304.

**Amore, Mario Daniele, Cedric Schneider, and Alminas Žaldokas**, "Credit supply and corporate innovation," *Journal of Financial Economics*, 2013, *109* (3), 835–855.

**Anderson, Ross J., Chris J. Barton, Rainer Böhme, Richard Clayton, Michel van Eeten, Michael Levi, Tyler W. Moore, and Stefan Savage**, "Measuring the Cost of Cybercrime," in "Workshop on the Economics of Information Security" 2012.

**Azoulay, Pierre, Christian Fons-Rosen, and Joshua S Graff Zivin**, "Does science advance one funeral at a time?," *American Economic Review*, 2019, *109* (8), 2889–2920.

**Blundell, Richard, Rachel Griffith, and John Van Reenen**, "Market share, market value and innovation in a panel of British manufacturing firms," *The Review of Economic Studies*, 1999, *66* (3), 529–554.

**Canayaz, Mehmet, Ilja Kantorovitch, and Roxana Mihet**, "Consumer Privacy and Value of Consumer Data," Technical Report 22-68 2022.

**Cohn, Jonathan B, Zack Liu, and Malcolm I Wardlaw**, "Count (and count-like) data in finance," *Journal of Financial Economics*, 2022, *146* (2), 529–551.

**Cong, L.W., D. Xie, and L. Zhang**, "Knowledge Accumulation, Privacy, and Growth in a Data Economy," *Management Science*, 2021, *67* (10), 6480–6492.

**_ , W. Wei, D. Xie, and L. Zhang**, "Endogenous Growth under Multiple Uses of Data," *Journal of Economic Dynamics and Control*, 2022, *104395.*

**Correia, Sergio, Paulo Guimarães, and Tom Zylkin**, "Fast Poisson estimation with high-dimensional fixed effects," *The Stata Journal*, 2020, *20* (1), 95–115.

**Dass, Nishant, Vikram Nanda, and Steven Chong Xiao**, "Truncation bias corrections in patent data: Implications for recent research on innovation," *Journal of Corporate Finance*, 2017, *44*, 353–374.

**Ettredge, Michael, Feng Guo, and Yijun Li**, "Trade secrets and cyber security breaches," *Journal of Accounting and Public Policy*, 2018, *37* (6), 564–585.

**Ewens, Michael, Ryan Peters, and Sean Wang**, "Measuring Intangible Capital with Market Prices," *Working Paper*, 2020.

**Farboodi, M. and L. Veldkamp**, "A Growth Model of the Data Economy," Technical Report 28427 2021.

**Farboodi, Maryam, Roxana Mihet, Thomas Philippon, and Laura Veldkamp**, "Big Data and Firm Dynamics," *AER Papers and Proceedings*, May 2019, *109*, 38–42.

**Florackis, Chris, Christodoulos Louca, Roni Michaely, and Michael Weber**, "Cybersecurity risk," *The Review of Financial Studies*, 2023, *36* (1), 351–407.

**Freeman, R.B., B. Yang, and B. Zhang**, "Data Deepening and Nonbalanced Economic Growth," Technical Report 3894511 2021.

**Giczy, Alexander V, Nicholas A Pairolero, and Andrew A Toole**, "Identifying artificial intelligence (AI) invention: A novel AI patent dataset," *The Journal of Technology Transfer*, 2022, *47* (2), 476–505.

**Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg**, "The NBER patent citation data file: Lessons, insights and methodological tools," 2001.

**_ , Adam Jaffe, and Manuel Trajtenberg**, "Market value and patent citations," *RAND Journal of economics*, 2005, pp. 16–38.

**Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches**, "Econometric Models for Count Data with an Application to the Patents-R & D Relationship," *Econometrica*, 1984, *52* (4), 909–938.

**Hoberg, Gerard and Gordon Phillips**, "Text-based network industries and endogenous product differentiation," *Journal of Political Economy*, 2016, *124* (5), 1423–1465.

**Hou, Y., J. Huang, D. Xie, and W. Zhou**, "The Limits to Growth in the Data Economy: How Data Storage Constraint Threats," Technical Report 4099544 2022.

**Howell, Sabrina T**, "Financing innovation: Evidence from R&D grants," *American Economic Review*, 2017, *107* (4), 1136–64.

**Jamilov, Rustam, Hélène Rey, and Ahmed Tahoun**, "The anatomy of cyber risk," Technical Report, National Bureau of Economic Research 2021.

**Jiang, Hao, Naveen Khanna, Qian Yang, and Jiayu Zhou**, "The cyber risk premium," *Available at SSRN 3637142*, 2020.

**Jones, C.I. and C. Tonetti**, "Nonrivalry and the Economics of Data," *American Economic Review*, 2020, *110* (9), 2819–2858.

**Kamiya, Shinichi, Jun-Koo Kang, Jungmin Kim, Andreas Milidonis, and René M Stulz**, "Risk management, firm reputation, and the impact of successful

cyberattacks on target firms," *Journal of Financial Economics*, 2021, *139* (3), 719–749.

**Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman**, "Technological innovation, resource allocation, and growth," *The Quarterly Journal of Economics*, 2017, *132* (2), 665–712.

**Lerner, Josh and Amit Seru**, "The use and misuse of patent data: Issues for finance and beyond," *The Review of Financial Studies*, 2022, *35* (6), 2667–2704.

**Silva, JMC Santos and Silvana Tenreyro**, "The log of gravity," *The Review of Economics and statistics*, 2006, *88* (4), 641–658.

_ **and** _ , "Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator," *Economics Letters*, 2011, *112* (2), 220–222.

**Wooldridge, Jeffrey M**, "Distribution-free estimation of some nonlinear panel data models," *Journal of Econometrics*, 1999, *90* (1), 77–97.

**Xie, D. and L. Zhang**, "Endogenous Growth with Data Generated During Production," Technical Report 4033576 2022.

# Appendix A Theoretical derivations

## A.1 Pervasive Cybercrime Impact

In the main text, a cyber attack was interpreted as any activity provoking a partial loss of the amount of data held by firm $i$ at date $t$. Specifically, under such interpretation, the criminal activity lowers the endowment of data of the firm from $z_i + \Delta_{i,t}\delta_{i,t}$ to $(1-\vartheta)(z_i + \Delta_{i,t}\delta_{i,t})$, with $\vartheta \in (0,1)$. Consequences of this assumption were thoroughly discussed. Essentially it led to a result such that the firm's stock of knowledge, its output, and its profits fall if the firm is under attack, with this effect being mitigated if the producer adopts some sort of cyber protection. The question that we ask in this additional note is whether this result is maintained or not if, instead of assuming the impact of crime over available data, one assumes that cybercrime has a pervasive penalizing effect over the firm's stock of knowledge. The justification for this new assumption finds support in the reasoning that it is not only current data that may be compromised by the attack, but the entire stock of knowledge that the firm has accumulated from past uses of data till the current period.

To implement the new assumption, one recovers the constraint that characterizes the time evolution of $\Omega_{i,t}$, letting, in this case, cybercrime to have an overall effect over the accumulation of knowledge, i.e.,

$$\Omega_{i,t+1} = (1-\vartheta)\left\{\left[\rho^2(\Omega_{i,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + (z_i + \Delta_{i,t}\delta_{i,t})\,\sigma_\epsilon^{-2}\right\} \tag{40}$$

The change in the formulation of the impact of cyber attacks generates new equilibrium outcomes. However, as explained below, it does not compromise the main qualitative results one has derived for the case in which crime only affects the data endowment of the time period under consideration.

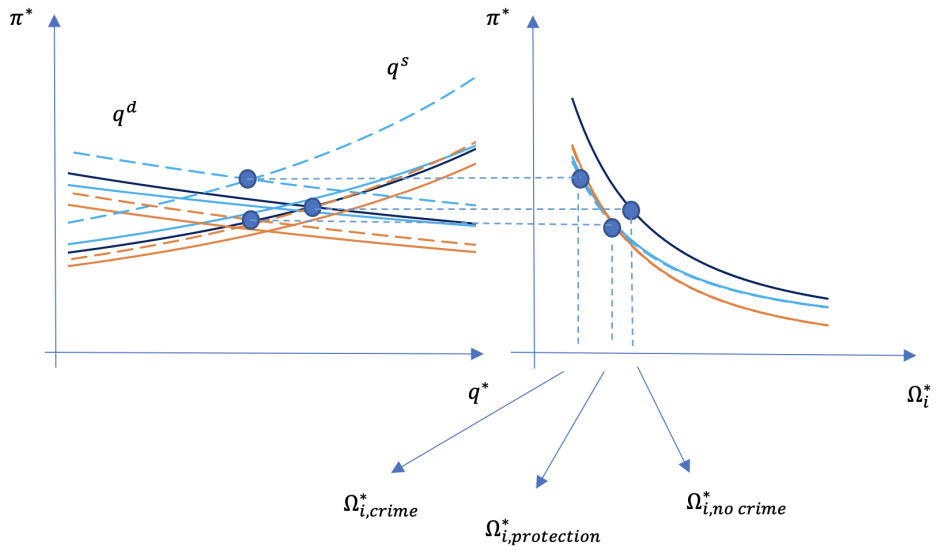
The optimality condition now writes as:

$$\left[\rho + \frac{\sigma_\theta^2}{\rho}(\Omega_{i,t} + \sigma_a^{-2})\right]^{-2} \frac{\pi_t \sigma_\epsilon^2}{\Delta_{i,t}} = \frac{\pi_{t-1}\sigma_\epsilon^2}{(1-\vartheta)\beta\Delta_{i,t-1}} - \left(\Omega_{i,t} + \Omega_{(l,h)}\right)^{-2} \tag{41}$$

Confronting (41) with the corresponding expression in the alternative case, one verifies that the only change is that term $1-\vartheta$ is no longer present in the l.h.s. of the equation. This will lead to a change in the analytical expression of the price (the steady state of (41) solved with respect to $\pi$) and also in the analytical expression of

the quantity of traded data (the steady state of (40) solved with respect to $\delta$). These changes also modify the position of the demand and supply curves of data leading to a new point of intersection, i.e., to new equilibrium levels of price and traded data. These perturbations in the equilibrium of the data market have no significant effects on the knowledge curves (curves in the knowledge-price space, which remain basically in the same location as in the benchmark case of crime influencing only current data).

Figure 13: Steady state equilibrium with pervasive cyber-crime




Legend: With cyber-protection, the stocks of knowledge and the aggregate output level are at an intermediate level between no crime and crime with no protection. Protection mitigates the negative impact of cybercrime, thus placing firms in a better position to use their data and create value; however, protection has a direct cost that hampers the stock of knowledge the firm can accumulate and, therefore, reduces the value of output as well. The market equilibrium analysis reveals that cyber-protection shifts both the demand curve and the supply curve down (for both trading positions of the two groups of producers), leading to the formation of a steady state with a lower equilibrium price.

The results are depicted in Fig. 13 for the same numerical example used in the text. The dashed lines represent demand and supply under the new specification. Both under cybercrime and cyber security, equilibrium results shift to points of higher price and lower traded data; however, when one looks at the graphic on the right, one observes that neither the knowledge curve under cybercrime or the knowledge curve under protection suffer visible changes (the dashed lines practically overlap the original curves). Consequently, the main result is identical to the one in the initial formulation: cybercrime disturbs the no crime equilibrium and cyber protection disturbs the crime

equilibrium in such a way that the stock of knowledge under cyber security remains somewhere in the middle between the cases of absence of crime and crime with no security.

As in the original case, the graphical analysis is undertaken for the scenario in which high data-intensity firms are sellers of data and low data-intensity firms are data buyers. The inverse scenario delivers similar qualitative results.