Data Innovation Complementarity and Firm Growth*

Anastassia Fedyk † Orlando Gomes ‡ Roxana Mihet §
Kumar Rishabh ¶

December 28, 2024

Abstract

This paper examines how complementarity between a firm's general innovation and data-security innovation affects firm outcomes in the modern data economy. Our theoretical model shows that when the importance of data and its protection increases, firms with high complementarity between data-security and non-data-security innovation enter a virtuous cycle. They take advantage of this complementarity to improve their broader predictive capabilities and extend their productivity frontier. Empirically, we propose a novel firm-level measure of data innovation complementarity based on the intersection of patent inventors who work on both data-security-related and non-data-security related patents. Leveraging the staggered introduction of Data Breach Notification Laws (DBNLs) across U.S. states as a quasi-exogenous shock, we provide robust empirical evidence that heightened incentives to protect in-house generated data activates this feedback loop. We find that firms with complementary data innovation processes experience significant increases in (overall) innovation and profitability, by not only enhancing their in-house data security measures but also integrating these innovations across other domains. In contrast, firms without complementary data experience negative effects from the data protection laws. Our results highlight the dual role of data in driving firm-level market power and innovation dynamics.

Keywords: Data economy, data complementarity, data feedback loop, growth, innovation.

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

*First version: January 31, 2023. This version: December 28, 2024. This paper previously circulated under the title "Data Risk, Firm Growth, and Innovation". We thank our discussants, Fabrice Collard, Steven Ongena, Luca Sandrini, Chi-Yang Tsou, and Baozhong Yang for their helpful feedback, as well as participants at FOFI 2024, TSE Digital Economics 2024, WEIS 2024, GRETA 2024, EEA 2024, SFI Annual Days 2023, the Economics of ICT 2023, Boca-ECGI 2023, EEA 2023, ERMAS 2023, and seminars at the University of Zürich, ETH Risk Center, Boston University, University of Lausanne, and University of Neuchâtel for useful feedback. We are also grateful for insightful suggestions from Tania Babina, Francesco Celentano, Andreas Fuster, Leonardo Gambacorta, Luise Eisfeld, Gerard Hoberg, Tarun Ramadorai, Norman Schürhoff, Laura Veldkamp, and Venky Venakteswaran. Roxana Mihet acknowledges generous funding for this project from The Sandoz Family Foundation - Monique de Meuron Programme, as well as from UNIL's Dean's Office. All authors declare no conflicts of interest related to this project. All errors are our own.

†Haas School of Business at UC Berkeley, fedyk@berkeley.edu.

‡ISCAL, omgomes@iscal.ipl.pt.

§University of Lausanne and SFI, roxana.mihet@unil.ch.

¶University of Lausanne, and University of Basel, kumar.rishabh@unil.ch.

1 Motivation

Data is transforming the modern economy (Farboodi et al., 2019). Data assets are becoming increasingly critical for modern firms, serving as direct inputs into production and as a way to glean valuable insight for better forecasting (Farboodi and Veldkamp, 2021). Protection of data assets is therefore an area of growing focus for data-intensive firms. At the same time, firms with superior data processing technology—for example, firms investing in artificial intelligence (AI) technologies—are reaping great benefits from their data stock, leading to significant growth. Much of this growth comes from increased innovation and new product creation (Babina et al., 2024a). Our paper is the first to directly measure firm-level complementarity between data-related innovation and other innovation activity, which we term "data complementarity." Using our novel measure, we document positive *data feedback loops*: when the importance of protecting data assets increases, firms with high data complementarity experience a virtuous cycle, whereby their increased data-security-related innovation spills over into other forms of innovation, leading to overall growth.

This paper examines the premise that firms with close links between data-related and non-data-related innovation derive a dual benefit from their data expertise and activities. First, they are better equipped to protect their valuable data assets, coming up with data-security-related innovations. Second, in firms with high complementarity, the same engineers and inventors who develop cutting-edge data protection systems are also relevant for other product innovation. With increased impetus to innovate in the data security space, these inventors create cross-domain spillovers that amplify the firms' competitive advantage. A striking example of this dynamic is Amazon: the firm's patented data-protection technology, originally developed to securely transmit financial information, became the foundation for their groundbreaking "1-Click ordering" system—one of Amazon's most cited patents and a hallmark of their operational efficiency and customer experience innovation.

To test this hypothesis, we develop a new measure of complementarity between a firm's innovation in the data security space and in other areas, which we term "data complementarity." The measure is based on patenting activity—looking at the intersection of inventors who work on data-security-related and non-data-security related patents. Specifically, we construct a list of data-security specialists, defined as inventors with at least 10% of their patent portfolio in data-security innovations, and measure complementarity through their participation in non-data-security patent teams. This team-level approach captures active integration of data-security expertise into broader innovation efforts.

A growing literature leverages patents to measure innovation in technologies such

as AI (Alderucci et al., 2020), firm linkages and specialization (Bena et al., 2021), and creative destruction (Kakhbod et al., 2024). Our paper is the first to leverage the rich patent data to measure complementarity between firms' data-related innovation and broader innovation activities.

Aside from resolving the challenge of measuring the interplay between firms' data-related and non-data-related innovation, we need to address a second empirical challenge: measuring variation in the value and importance of data assets. The value of data differs between firms and fluctuates over time, but these differences are difficult to identify empirically and likely to be endogenous to firms' activities. In order to examine the potential data feedback loop, we need exogenous shocks to the salience of data.

We circumvent this identification and measurement challenge by leveraging the staggered adoption of Data Breach Notification Laws (DBNLs) across U.S. states, which exogenously heightened the salience of data for firms. DBNLs mandate organizations to notify individuals, regulatory authorities, and other stakeholders of security breaches involving unauthorized access, disclosure, or loss of personal data. Our strategy is to compare the financial decisions and innovation activities of firms located in early-treated states to those of firms located in late-treated states, taking into account the 'forbidden comparison' in staggered difference-in-difference models (Goodman-Bacon (2021)). This approach allows us to leverage the exogenous variation provided by the staggered implementation of DBNLs across states to examine the causal effects of data salience. While these laws impose costs and operational challenges on firms (Boasiako and Keefe, 2021; Liu and Ni, 2023), they also spur firms to reassess and enhance their data strategies. We show that high data-complementarity firms respond to these shocks by innovating more—particularly in areas that integrate their data expertise into broader applications. In contrast, low data-complementarity firms struggle to adapt to the rising prominence of data, widening the performance gap between the two groups.

Our empirical analysis reveals three key insights. First, high data-complementarity firms exhibit significantly higher innovation output, in both process and product patents compared to low data-complementarity firms in response to increased data salience brought on by the DBNLs. Second, this increased innovation translates into a boost in profitability, higher market share growth, and greater differentiation from competitors. Third, we highlight a mechanism of cross-pollination within high data-complementarity firms: the expertise developed to manage and process data also drives advancements in other innovations, with increased self-citations back to the data-security-related patents, leading to a positive data feedback loop.

In terms of innovation output, for high data-complementarity firms there is an observed increase of approximately 20 citation-weighted patent counts two to five years after the passage of the DBN laws, which corresponds to a 115.6% increase in citation-

weighted patent counts relative to the mean. In contrast, low data-complementarity firms experience a slight decline of approximately 1 to 2 citation-weighted patent counts in the same post-DBN laws period, which corresponds to a 11.6% decline in citation-weighted patent counts relative to the mean. In terms of standard deviations, the post-adoption effect of data breach notification laws corresponds to an increase of approximately 0.20 standard deviations for high-complementarity firms and a decrease of approximately 0.02 standard deviations for other firms. The increased patenting by high data-complementarity firms in response to the DBN laws comes in the form of both increased process patents and increase process patents.

As direct evidence of the knowledge spillover mechanism (whereby high data-complementarity firms benefit from their data-related inventors innovating more in the data security space in response to the regulation, and then transferring this knowledge into other spheres), we look at citations on data-security-related patents. High data-complementarity firms increase citations of data-security-related patents in general and their own data-security-related patents (self-citations) in particular, after the passage of DBN laws. This showcases the importance of the data-security innovation—prompted by the heightened costs of data breaches due to DBN laws—for the high data-complementarity firms' broader innovation activities.

Does the positive effect on innovation for high data-complementarity firms translate into operational gains? We observe that these firms experience a large increase in profitability after the passage of DBN laws, reaching as much as 4% additional ROA (approximately one third of the inter-quartile range in ROA across the firms in our sample). High data-complementarity firms also increase their market shares and experience decreases in their Hoberg and Phillips (2024) fluidity measures. Thus, these firms face reduced threats from competitors and are able to gain ground as a result of the innovation they perform in response to the data security laws.

In the second part of the paper, we develop a theoretical framework to conceptualize our empirical findings on the data feedback loop—the interaction between firms' data complementarity measures and digital innovation. We construct a heterogeneous-firm growth model where data optimizes business processes, but is vulnerable to loss and destruction. Indeed, recent work (Scherbina and Schlusche (2022)) highlights that firms face significant vulnerabilities from data destruction, in addition to the well-studied threat of data theft. Firms vary in the complementarity between their data-related and non-data-related innovation and can protect themselves against data loss. High data-complementarity firms develop tailored in-house security solutions, enhancing the quality of their products. Low data complementarity firms purchase these non-rival data security solutions from high data-complementarity firms. Both types of firms benefit from mitigating this vulnerability: low data-complementarity firms preserve

data, while high data-complementarity firms gain additional innovation spillovers that improve product quality. This model allows us to conduct counterfactual analyses to explore these dynamics further.

Our research is nestled in the growing empirical and theoretical literature on data as a critical asset in a firm, illustrating its dual role as a source of risk (Mihet and Philippon (2019)), but also a catalyst for innovation (Babina et al. (2024a)). Data has to potential to propel firms toward growth and competitive advantage (Crouzet and Eberly, 2019, 2021), yet as a critical asset it also poses significant challenges in terms of potential data loss or breaches. Our work contributes to the understanding of these dual facets by examining how increased salience of data vulnerabilities can contribute to innovation in certain firms because of complementarity between data-related and non-data-related innovation activities, leading to a positive data feedback loop. Our results bring a novel and nuanced view that goes beyond the classic focus on firm valuation and equity returns impacted by data risk (Jamilov et al., 2021).

In doing so, we also contribute to the literature on innovation fueled by data and related technologies. For example, Babina et al. (2024a) find that the primary effect of artificial intelligence (AI) technology is firm growth fueled by increased innovation, suggesting that new data-intensive technologies are contributing to general product innovation. We offer a novel angle to this discussion by directly speaking to the complementarity between data-related and non-data-related innovation. As the salience of data and associated breaches increases, firms with low data innovation complementarity experience adverse effects, but firms with high preexisting complementarity between data-related and non-data-related innovation are able to not only mitigate the risks, but also leverage them as a springboard for broader innovation. This dual focus on threat and opportunity aligns with recent work on the systemic significance of data and AI in sectors like finance and technology (Duffie and Younger (2019); Aldasoro et al. (2022), Akey et al. (2018)), as well as insights into firms' strategic adaptation to technological disruptions (Jiang et al. (2021)).

Finally, our paper relates to the growing debate of the valuation of data (Goldstein et al., 2024; Farboodi et al., 2022) and who should own data (Jones and Tonetti, 2020; Babina et al., 2024b). Data is a critically important asset in the modern economy, and its protection is important for both consumer welfare and firm performance. Our results highlight a novel effect of customer data protection laws: they encourage firms to innovate, and those firms that have well-integrated data-related and non-data-related innovation teams are able to reap unexpected benefits in the form of increased innovation in other spheres. On the one hand, this is a positive effect, spurring innovation and profitability improvements. On the other hand, it leads to increased market dominance of high data-complementarity firms, which is a potential dark side. Thus, our

results highlight the importance of balancing data protection mandates with innovation incentives in optimal regulatory frameworks.

The remainder of the paper proceeds as follows. Section 2 describes the data and the construction of our key empirical measures. Section 3 details our staggered difference-in-differences strategy around the passage of state-level DBN laws. Section 4 presents the empirical results. Section 5 develops a theoretical model of the data economy with data risk and protection and conducts comparative statics exercises. Section 6 concludes with policy implications and avenues for future research.

2 Data and Measures

We introduce our novel measure of data complementarity based on patent data. After that, we describe the other datasets and measures used in the paper: overall and data-security-related innovation measured using several complementary datasets, and firm financial and operational measures from CRSP-Compustat.

2.1 Data Complementarity Measure

To study the economic effects of data feedback loops, we introduce a novel method for identifying high data-complementarity firms using publicly accessible patent data. This approach, which surpasses simple measures of a firm's data stock, is the first to assess the interplay between data-related and broader innovation activities. It provides a replicable framework that can be extended across time, geographies, and both public and private firms. Our index captures the direct and spillover value of data within a firm's operations, reflecting its critical economic importance as an enabler of innovation and productivity across diverse domains.

We construct the new complementarity measure in the following steps: (i) The first step is to construct a list of data-security engineers and innovators at a given firm i, using USPTO patent data. A patent inventor is included in this list if at least 10% of their patent portfolio consists of data-security patents. This threshold ensures meaningful specialization in data-security innovation rather than incidental involvement. (ii) The second step is to calculate, for each patent p assigned to the firm i, the share of data-security inventors from the previous list in patent p's inventor team, capturing data-complementarity expertise embedded in each innovation project. (iii) The third step is to aggregate the patent-level complementarity measure to the firm-year level. For each firm-year, we compute the average share of data-security inventors across all *non-data security patents* over the preceding five-year window. This rolling window captures recent innovation activity while avoiding look-ahead bias. (iv) Finally, we

classify a firm i in year t as a high data-complementarity firm if its mean share of data-security inventors in non-data-security patents equals or exceeds the 75th percentile across the panel. This classification is sticky – once attained, a firm retains its high data-complementarity status in subsequent years.

This measure of data complementarity based on the patent inventor team has several key features that make it a robust and economically meaningful indicator of complementarity. Looking at patent inventors, who are rare and significant drivers of firm value, leads to a very targeted measure. By capturing instances where firms allocate scarce data-security expertise to non-data-security projects, the measure reflects revealed preference in resource allocation—a signal of true complementarity rather than organizational coincidence. The team-level approach ensures precise identification of knowledge transfer, as it captures actual collaborative instances where data-security expertise directly contributes to non-data-security innovation rather than just potential for such transfer. Human capital is a critically important input into technology use, and cross-sharing of complementary expertise from one area of the firm can substantially improve performance in another area (Fedyk et al., 2023). The rolling five-year window with a sticky high-complementarity designation further by conceptually focusing on structural rather than transitory complementarity.

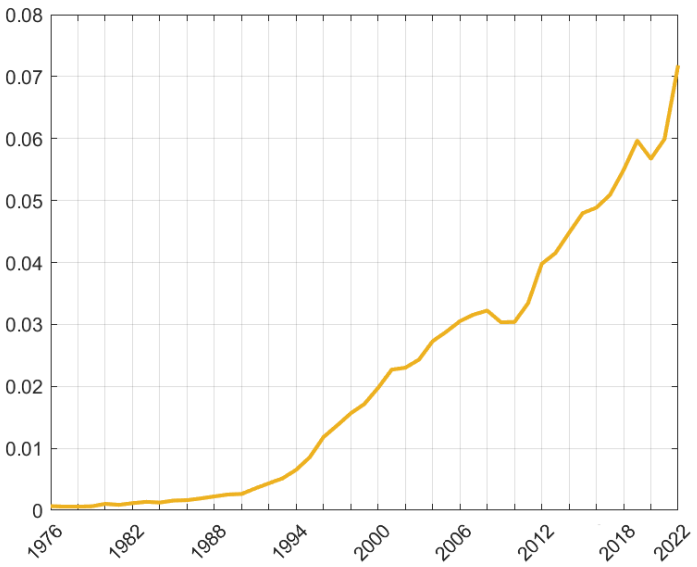
2.2 Innovation

Broad Innovation. Firms' patent activity captures their innovation output. Following the literature on innovation, we assess patents filed by the firms by taking into account their scientific value (Kogan et al., 2017; Aghion et al., 2013; Howell, 2017). We count the number of patents filed, weighted by the number of forward citations they receive. The idea is that the more scientifically important a patent is, the more citations it receives (Hall et al., 2005; Kogan et al., 2017). Following best practices in the literature, we adjust the count for truncation bias. As the citations occur over time, a simple count of citations underestimates the importance of the patents that were issued towards the end of our sample period (Lerner and Seru, 2022; Dass et al., 2017). We correct for this issue using the methodology proposed by Hall et al. (2001). All our patent data come from the publicly available database maintained by Kogan et al. (2017).

Data-security-related innovation. We measure data-security-related innovation using the citation-weighted count of cyber-security patents filed by a given firm in a given year. A patent is classified as a cyber-security patent if the United States Patent and Trademark Office (USPTO) assigns that patent at least one Cooperative Patent Classification (CPC) code associated with cyber security. For instance, CPC

code G06F21 is titled "Security arrangements for protecting computers, components thereof, programs or data against unauthorised activity." The time series of our aggregate cyber-security patent measure (computed across patents of all firms), depicted in Figure 1, indicates consistent growth in data security innovation over time. By 2022, the fraction of cyber security patents reaches seven percent of all patent filings.

Figure 1: Share of data security (i.e., cyber security) patents among all patents filed



Legend: This figure shows the proportion of data security (i.e., cyber security) patents out of all patents filed in a given year from 1976 to 2022. A cyber security patent is defined as a patent to which the USPTO assigns at least one CPC code pertaining to cyber security.

2.3 Financial Data

We obtain firm level financial information from the merged CRSP-Compustat database. We calculate the following financial variables and ratios to use as control variables in our baseline regressions: 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 variables at 0.5% and 99.5% of the corresponding distribution.

Table 1 presents summary statistics on the firm variables of interest. Innovation activity is quite skewed, with more than 50 percent of firm-year observations not recording any positive patent activity.

Table 1: Descriptive statistics

	N	Mean	SD	p10	p25	p50	p75	p90	p99
I. Innovation									
Patents filed: c-wtd count	60007	17.32	98.34	0	0	0	0	12.05	546.87
Cyber security patents: c-wtd count	60007	0.98	8.16	0	0	0	0	0	29.16
Non-cyber security patents: c-wtd count	60007	15.65	89.02	0	0	0	0	11.08	474.77
Product patents: c-wtd count	60007	9.45	54.95	0	0	0	0	5.53	294.85
Process patents: c-wtd count	60007	4.59	28.44	0	0	0	0	2.39	144.03
Share of product patents in c-wtd count	11621	0.65	0.34	0	0.43	0.73	1	1	1
R&D expenditure	60007	63.50	308.21	0	0	0	15.96	84.80	1549.91
II. Financial attributes									
Total assets (log)	60007	6.73	2.17	3.79	5.23	6.80	8.20	9.51	11.93
Tobin's Q	59896	2.04	1.93	0.94	1.04	1.39	2.19	3.79	11.16
Tangibility	57659	0.20	0.24	0.01	0.02	0.10	0.29	0.63	0.90
Return on assets	57511	0.00	0.30	-0.26	0.01	0.07	0.13	0.20	0.43
Book-to-market ratio	59896	0.63	0.79	0.09	0.25	0.50	0.86	1.30	3.97
Cash-to-asset ratio	60006	0.21	0.25	0.01	0.03	0.10	0.29	0.62	0.97
Leverage	59764	0.24	0.24	0	0.04	0.18	0.38	0.57	1.04

Legend: N refers to the total number of firm-year observations. The citation-weighted (c-wtd) patent counts weigh each patent with the forward citation the patent receives, adjusting for the filing vintage. p10-p99 refer to the 10th to 99th percentile values.

3 Empirical Strategy: Data Breach Notification Laws

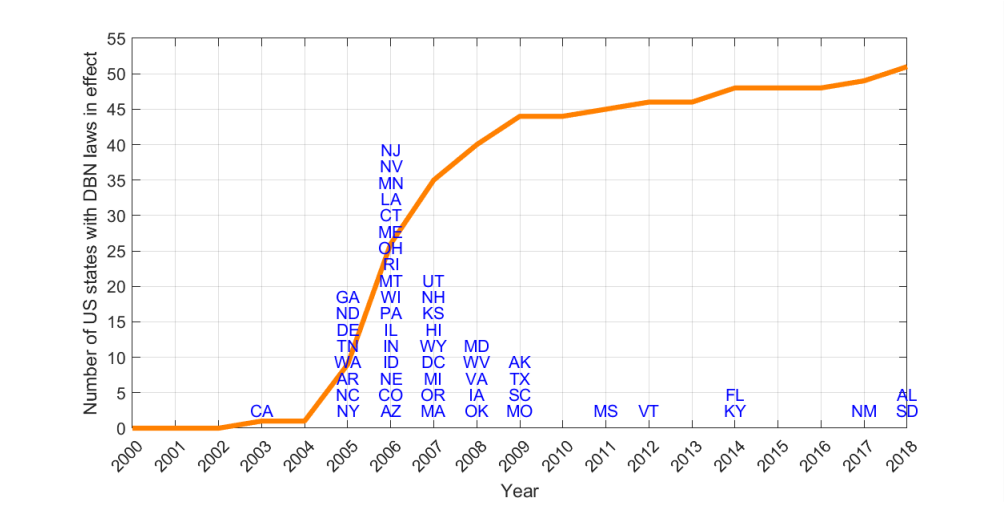
In order to estimate the causal impact of data importance and salience on firm performance, we exploit the staggered implementation of Data Breach Notification Laws (DBNLs) across U.S. states, which created an exogenous increase in the importance of data for firms.

DBNLs require organizations to notify individuals, regulators, and other stakeholders of security breaches involving unauthorized data access, disclosure, or loss. Our analysis compares the financial and innovation decisions of firms in states that adopted DBNLs earlier to the decisions of firms in states that implemented them later, addressing the 'forbidden comparison' issue in staggered difference-in-difference models (Goodman-Bacon (2021)). This approach utilizes the exogenous variation from the staggered rollout of DBNLs to identify the causal effects of heightened data salience. Although these laws introduce compliance costs and operational challenges for firms (Boasiako and Keefe, 2021; Liu and Ni, 2023), they also prompt firms to reevaluate and strengthen their data management strategies.[1] All 50 states have enacted their own versions of DBNLs, starting in 2003 with California and ending in 2018 with Alabama

[1]Data Breach Notification Laws (DBNLs) in the United States mandate firms to inform individuals affected by a data breach that involves their personal information. Typically, these laws require companies that experience a data breach to notify affected individuals within a specified time-frame, often ranging from 30 to 90 days after the breach is discovered. The notification usually includes details about the nature of the breach, the type of information compromised, and the steps individuals can take to protect themselves. In addition, some states require organizations to notify state authorities or consumer reporting agencies depending on the scale and severity of the breach. The laws also have provisions outlining penalties for non-compliance, aiming to hold organizations accountable for safeguarding individuals' personal data.

and South Dakota. By 2008, more than half of the states had adopted a DBN law, as shown in Figure 2.

Figure 2: State adoption of DBNLs



Legend: This figure reports the first time that each state enacts a data breach notification law specifically containing data security breach notification provisions. For example, Nevada introduced a data breach statute in January 2005, but it only required notification provisions for general data provisions in January 2006; thus, in our sample, Nevada appears as a 2006 adoption of DBN law. The source for the data on the timing of the DBNLs is Perkins Coie LLP (2023).

DBNLs are appropriate instruments to address measurement and endogeneity issues for quantifying the economic importance of data-feedback loops for several reasons: (i) The staggered implementation of DBNLs across states is driven by legislative processes and is unlikely to be correlated with individual firms' prior levels of innovation or specific data-complementarity profiles (as we confirm empirically with our pre-trend analysis). The staggered roll-out introduces exogenous variation in data salience that is not directly influenced by firms' endogenous decisions, providing a robust source of exogenous variation; (ii) DBNLs significantly increase the cost and consequences of data breaches for firms, thereby directly affecting the salience of data assets for firms' decision-making. This is evident from the increased efforts by firms to enhance data protection measures following the adoption of these laws, as documented in previous studies (Boasiako and Keefe, 2021; Liu and Ni, 2023; Huang and Wang, 2021); and (iii) The legal requirement to disclose data breaches under DBNLs ensures that firms cannot under-report or hide incidents, leading to an accurate and consistent measure of data salience across states and over time.

Our empirical strategy explores the staggered implementation of Data Breach Notification Laws (DBNLs) in the United States, which increased firms' importance of guarding their data. This scenario is analogous to the fintech disruption discussed by

Jiang et al. (2021), where firms' ability to innovate and adapt to regulatory changes is crucial for their survival and growth. Similarly, we examine how different firms—based on their pre-existing complementarity between data-related and non-data-related innovation—respond to increased data salience by innovating in data security and allocating resources to manage business operations effectively. We compare the innovation activities of firm headquarters located in early-treated states to those of firm headquarters located in late-treated states.

While DBNLs provide a powerful identification strategy, there are potential limitations to consider. Firms operating in multiple states may be affected by DBNLs earlier than the state of their headquarters, potentially leading to spillover effects via the firm's internal policies, practices, or economic activities, potentially contaminating the control group. This could attenuate the estimated impact of DBNLs. However, these concerns work in our favor, as the estimation provides a lower bound for the impact of DBNLs on treated firms relative to a control group that may have already reacted beforehand. Additionally, differences in the stringency and enforcement of DBNLs across states could lead to heterogeneous effects. While we control for state fixed effects and perform robustness checks, some variation in the impact of DBNLs may still exist. Furthermore, the nature of data salience and firms' responses to DBNLs may evolve over time. Our analysis accounts for these dynamics by including time fixed effects and examining the effects over different time horizons.

While no instrument is perfect, we believe we adequately capture firm responses to an increase in data salience by estimating lower bounds. Ideally, these state laws would have affected only firms in the respective state, allowing us to estimate the true average effect rather than the lower bound. In Appendix B, we discuss more extensively why Data Breach Notification Laws (DBNLs) are appropriate instruments, their potential limitations, and the measures we have taken to account for any potential concerns.

We also make sure we take into account the latest critiques in the literature on staggered difference-in-difference estimation. A very recent literature (Baker et al., 2022; Goodman-Bacon, 2021) has uncovered two vital econometric issues in standard staggered difference-in-difference methods such as linear two-way fixed effects (henceforth, TWFE): (1) there is a possibility of bias due to 'forbidden comparisons', and (2) there is a possibility of bias and/or inefficiency due to mis-specification in the presence of right-skewed dependent variables. The first issue refers to the potential for standard dynamic two-way fixed effects methods to suffer from an aggregation problem of treatment effects over some valid comparisons but also over some 'forbidden comparisons'. Specifically, TWFE compare already-treated units (as controls) with the later-treated units (as treated). When the treatment effects are heterogeneous over time or across treatment units, this may lead to biased average treatment effects in the

treated (ATT) estimates. The second issue of mis-specification in the presence of right-skewed dependent variables is problematic, because using a $log(1+y)$ transformation of the dependent variable, a log-linear, or an inverse hyperbolic sine (IHS) regression produces inconsistent and biased estimates. Another method to reduce skewness, the negative binomial regression, does not work with fixed effects. This leaves us with three models that admit fixed effects and produce unbiased estimates: linear, Poisson, and rate regressions. Linear regressions can be admitted can avoid bias and inconsistency issues, despite producing high-variance estimates (Cohn et al., 2022). This will make it harder to obtain significant results, but at least the estimates will be unbiased and consistent with the correct sign. Positive significant results will suggest that *despite* the method producing high variance estimates, there is evidence data breach notification laws have an effect on firm financial and innovation activities.

In our analysis, we use the Borusyak et al. (2022) linear method (henceforth, BJS) to address the first challenge of 'forbidden comparisons'. The BJS method is unbiased and consistent (Cohn et al. (2022)), despite being inefficient. Other popular methods that account for 'forbidden comparisons' are Callaway and Sant'Anna (2021), Sun and Abraham (2021), and de Chaisemartin and d'Haultfœuille (2020), among others. The BJS estimator is the most efficient under the assumption of parallel trends because it uses all pre-treatment data in the estimation and it is robust to cases when treatment effects vary arbitrarily. The first estimation that we run is a linear difference-in-difference regression accounting for 'forbidden comparisons' using the BJS 3-step imputation representation for the efficient estimator, which proceeds as follows:

1. Within the untreated observations only, estimate the λ_i and δ_t (by $\hat{\lambda}_i^\star$, $\hat{\delta}_t^{\,\star}$) by OLS in equation (1), where λ_i is unit (i.e., firm) fixed effect, δ_t is year fixed effect:

$$Y_{it} = \lambda_i + \delta_t + \epsilon_{it}; \tag{1}$$

2. For each treated observation with $w_{it} \neq 0$, set $\hat{Y}_{it} = \hat{\lambda}_i^\star + \hat{\delta}_t^{\,\star}$ and $\hat{\tau}_{it}^\star = Y_{it} - \hat{Y}_{it}(0)$ to obtain the estimate of τ_{it};

3. Estimate the target τ_w by a weighted sum $\hat{\tau}_w^\star = \sum_{it} w_{it}\hat{\tau}_{it}^\star$;

This model allows us to estimate unbiased and consistent dynamic treatment effects using panel data, where Y_{it} is the year t outcome measure for firm i, $Y_{it}(0)$ is the year t stochastic potential outcome of firm i if it were never treated, $\Omega_1 = \{it \in \Omega | treated = 1\}$ is the set of treated observations (i.e., firms headquartered in a state that has adopted a DBNL), $\Omega_0 = \{it \in \Omega | treated = 0\}$ is the set of untreated (i.e., never-treated and not-yet-treated) observations, $\tau_{it} = \mathbb{E}[Y_{it} - Y_{it}(0)]$ represents the causal effects on the treated

observations $it \in \Omega_1$, and w_{it} are BJS-derived pre-specified non-stochastic weights that depend on treatment assignment and timing, but not on realized outcomes.

4 Empirical Results on The Data Feedback Loop

In this section, we explore the impact of Data Breach Notification (DBN) laws on firm innovation activities, as proxied by their patenting activity, as well as on firm profitability and market power. The results show different patterns for firms with low versus high data-complementarity, where the DBNLs impose a cost on the former but create an opportunity for the latter. We end by discussing the mechanism behind this result: the expertise spillovers from data-security-related to non-data-security related innovation.

4.1 Innovation and Profitability

Data breach notification (DBN) laws could impact firm innovation by imposing stricter security requirements and compliance costs, which divert resources away from innovation and discourage risky, data-driven projects. At the same time, these laws incentivize firms to innovate in security and data management technologies to mitigate business threats and maintain trust. The overall effect depends on whether the regulatory burden outweighs the opportunities for innovation created by the need for better data protection practices.

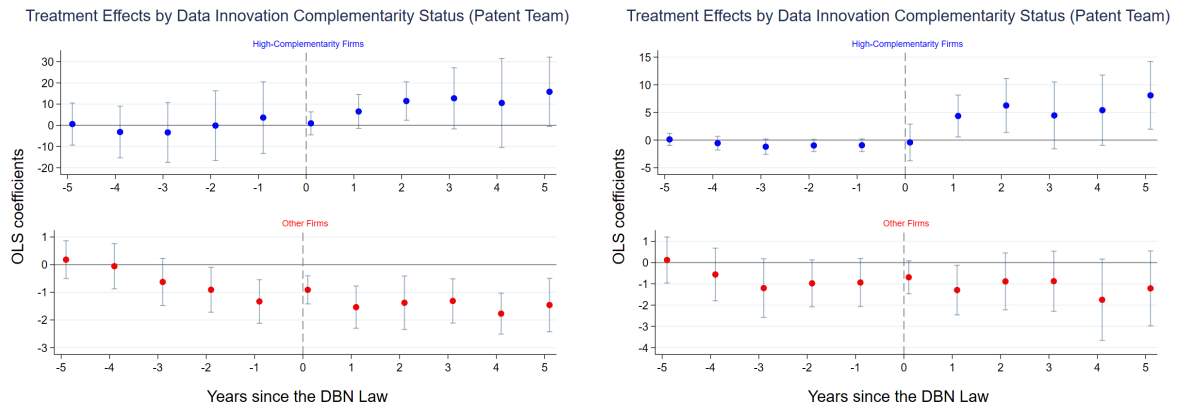
Figure 3 presents the BJS-weighted dynamic heterogeneous treatment effects of citation-weighted patent counts, separately for high data-complementarity firms and low data-complementarity firms. The left-hand panel allows heterogeneous pre-trends, while the right-hand panel assumes common pre-trends for both groups but estimates the average treatment effects on the treated (ATT) separately post-treatment.

High data-complementarity firms (Top panels) experience an *increase* in overall innovation after the adoption of DBNLs, with effect sizes on the order of an additional 10 citation-weighted patents per year (more than half the mean of the outcome variable). On the other hand, low data-complementarity firms experience a drop in overall innovation with the adoption of DBNLs, although the result is insignificant when accounting for common pre-trends. The combination of the results in Figure 3 suggests that the adoption of the DBN laws imposes a cost on firms that may detract from non-data-related innovation, *but* high data-complementarity firms actually increase their overall innovation activities, consistent with potential expertise spillovers from their data-related innovation work to other forms of innovation within the firm.

In untabulated analysis, we examine the impact specifically on cyber-security

Figure 3: Citation-weighted Patent Count

Left: Heterogeneous pre-trends Right: Common pre-trends



This figure plots BJS dynamic heterogeneous treatment effects of citation-weighted patent counts around the staggered adoption of DBN laws ('0' event) across U.S. states. Effects are estimated separately for high data-complementarity firms (Top panels) and low data-complementarity firms (Bottom panels), under heterogeneous pre-trends (Left panels) and common pre-trends (Right panels) assumptions. A firm-year is classified as high-complementarity if its mean share of data-security specialists participating in non-data-security patent teams over the preceding five years exceeds the 75th percentile of the distribution; this classification is sticky (once attained, it remains in effect in subsequent years). Patent counts are citation-weighted as one plus the patent's citation stock (normalized by same-filing-year average) as of the observation date.

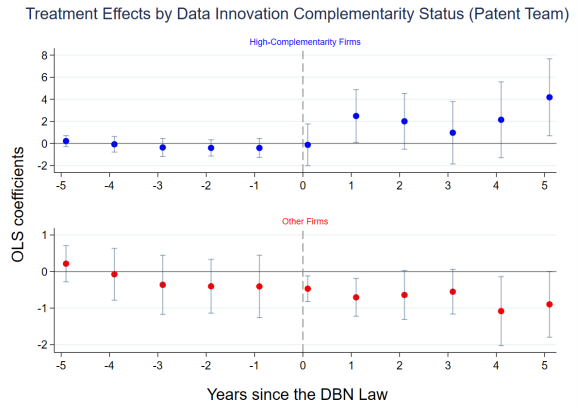
patents. Both high data-complementarity and low data-complementarity firms experience increases in their cyber security patents, although the effect for low data-complementairty is economically small and statistically insignificant. For high data-complementarity firms, DBNLs increase the citation-weighted count of cyber-security patents by up to 5 per year, roughly half the effect on overall patenting activity. Those, high data-complementarity firms are able to increase their cyber-security innovation more dramatically than low data-complementarity firms in response to the new legislature, and they leverage this increase in data-related innovation to also improve their non-data-related innovation capabilities.

We next examine differences in patent type (product vs. process patents) after the adoption of DBNLs. Product patents signify both the introduction of new products and enhancements in the quality of existing ones, while process patents capture improvements to organizational efficiency (Babina et al. (2024a)). To distinguish profuct versus process patents, we utilize the patent claims dataset provided by Ganglmair et al. (2022), which categorizes patent claims into product and process claims. We classify a patent as a product (process) patent if 50 percent or more of its claims are specified as product (process) claims, following the method described in Babina et al. (2024a).

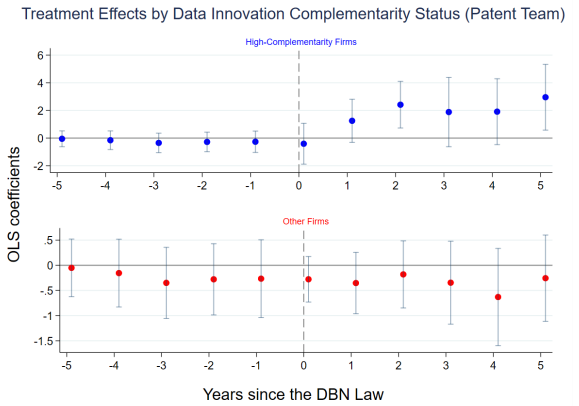
Figure 4 presents the BJS-weighted dynamic heterogeneous treatment effects of citation-weighted product patents count in the Left panels and citation-weighted pro-

Figure 4: Citation-weighted Patent Count: Product and Process Patents

Left: Product Patents



Right: Process Patents

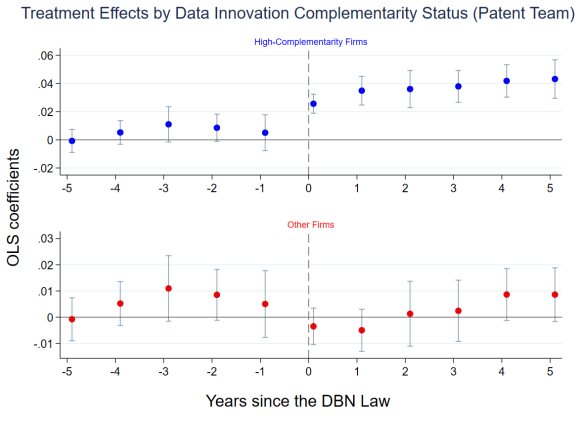


This figure plots BJS dynamic heterogeneous treatment effects of citation-weighted product and process patent counts around the staggered adoption of DBN laws ('0' event) across U.S. states. Effects are estimated separately for high data-complementarity firms (Top panels) and low data-complementarity firms (Bottom panels), under heterogeneous pre-trends assumptions, for citation-weighted product patents (Left panels) and process patents (Right panels). Patents are classified as product (process) patents if 50 percent or more of their claims are product (process) claims, based on the claims classification of Ganglmair et al. (2022). A firm-year is classified as high data-complementarity if its mean share of data-security specialists participating in non-data-security patent teams over the preceding five years exceeds the 75th percentile of the distribution; this classification is sticky (once achieved for a given firm, it remains in effect in subsequent years). Patent counts are citation-weighted as one plus the patent's citation stock (normalized by same-filing-year average) as of the observation date.

cess patent counts in the Right panels. The effects are computed separately for high data-complementarity firms (Top panels) and low data-complementarity firms (Bottom panels). The results show that high data-complementarity firms exhibit an increase in both process and product patents, and the effect on product patents, highlighting that these firms end up using the new regulation as a springboard for innovation, creating new products as a result of the need to better protect their data assets. In contrast, low data-complementarity firms experience an insignificant declines in both types of patents.

Changes in innovation, such as increased patenting activity, could translate into changes in profitability because they signal improvements in a firm's competitive edge. Patents represent novel technologies or processes that can enhance product quality, reduce production costs, or create new revenue streams. These innovations can enable firms to capture greater market share, command higher prices, or achieve cost savings, translating directly into improved profitability. As a result of the innovation, firm assets can be utilized more effectively, generating higher financial returns relative to their investment base. This is the reason we now turn our attention to the impact of Data Breach Notification (DBN) laws on firm profitability, proxied by return on assets (ROA).

Figure 5: Profitability (ROA)



This figure plots BJS dynamic heterogeneous treatment effects of firm profitability around the staggered adoption of DBN laws ('0' event) across U.S. states. Effects are estimated separately for high data-complementarity firms (Top) and low data-complementarity firms (Bottom), under heterogeneous pre-trends assumptions. Profitability is measured by Return on Assets (ROA), defined as the ratio of Operating Income Before Depreciation to Total Assets. A firm-year is classified as high-complementarity if its mean share of data-security specialists participating in non-data-security patent teams over the preceding five years exceeds the 75th percentile of the distribution; this classification is sticky in subsequent years.

Figure 5 shows that high data-complementarity firms (Top panel) exhibit a sustained, long-term increase in their profitability after the adoption of DBNLs, whereas low data-complementarity firms (Bottom panel) experience a null effect. The effect for high data-complemnetarity firms is highly significant and economically large: they experience an increase in profitability going up to more than 4% (approximately one third of the inter-quartile range of ROA in the data) a few years after the passage of DBNLs. Increased profitability (ROA) for high data-complementarity firms is evidence of a data-feedback loop, reflecting how these firms leverage their complementary expertise to turn potential costs (more stringent data breach reporting regulation) into opportunity: by innovating more, these firms are able to not only protect their data, but also create new products and process improvements. This leads to enhanced profitability. The virtuous cycle leads to compounding advantages over competitors who lack similar data capabilities.

Increased profitability is indicative of an impact on market power because it enables high data-complementarity firms to solidify their dominant positions. Higher returns allow these firms to invest in further data acquisition and technological improvements, creating barriers to entry for smaller competitors. Over time, this entrenched advantage enables high data-complementarity firms to exert greater control over pricing, market dynamics, and consumer behavior, reinforcing their influence within the market. We explore this dynamic futher in the next subsection.

4.2 Market Dominance of High Data-Complementarity Firms

The interplay between data-feedback loops and market power offers insights into how data protection regulations, such as Data Breach Notification (DBN) laws, can reinforce the dominance of high data-complementarity incumbents. The data-feedback loop—where enhanced data processing capabilities lead to greater (overall) innovation, competitive advantage, and improved market positioning—is particularly pronounced in firms that heavily invest in data technologies. DBNLs, while aimed at increasing data protection, raise compliance costs for firms with less integrated and complementary data innovation teams, while creating opportunities for high data-complementarity firms. This may inadvertently create barriers to entry and further consolidate the market power of high data-complementarity incumbents.

The mechanism is intuitive: compliance with DBNLs requires investments in data security infrastructure, which are less burdensome for high data-complementarity firms due to economies of scope as their specialized data engineers and inventors, while working to improve data security, can produce positive spillovers to the firm's other innovation activities. This dynamic makes it increasingly challenging for non-incumbents to compete effectively with these incumbent firms, thereby reducing the market fluidity experienced by these firms—a measure of how easily firms can enter and compete within a given industry space (Hoberg et al. (2014)).[2] In contrast, low data-complementarity firms are less able to turn the heightened data security costs into innovation opportunities, so they are less likely to experience reduced threats from competitors after the regulations.

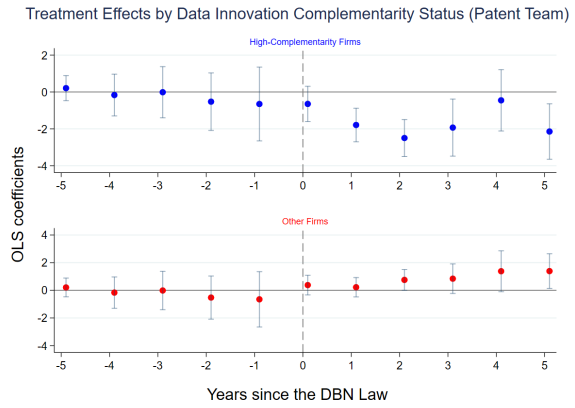
To test these hypotheses, we analyze the impact of DBNLs on changes in market shares and market fluidity. Changes in market shares are a proxy for market power, reflecting firms' ability to capture the market. We expect DBNLs to increase market shares for high data-complementarity firms due to their strengthened market positions post-regulation. At the same time, we anticipate a decrease in market fluidity for high data-complementarity firms, as their improved innovation and profitability reduces the threat of entry of new competitors.

Figure 6 illustrates the treatment effects of DBNLs on fluidity (Left panels) and changes in market share (Right panels), separately for high data-complementarity firms (Top panels) and low data-complementarity firms (Bottom panels). Our findings indicate that DBNLs boost the market share of firms with high data-complementarity. Mar-

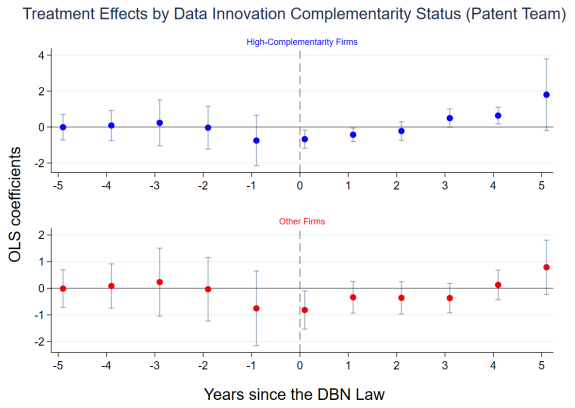
[2]Fluidity is derived from the similarity of firms' business descriptions in regulatory filings (e.g., 10-Ks) to the descriptions of their competitors, reflecting competitive overlap and potential market dynamics. High fluidity indicates a dynamic, competitive environment where new firms can easily challenge incumbents. Conversely, low fluidity suggests entrenched market power, where incumbents face limited competition due to barriers such as regulation, technology, or resource constraints.

Figure 6: Market Power Measures

Left: Self-fluidity



Right: Change in Market Share



This figure plots BJS dynamic heterogeneous treatment effects around the staggered adoption of DBN laws ('0' event) across U.S. states. Effects are estimated separately for high data-complementarity firms (Top) and low data-complementarity firms (Bottom), under heterogeneous pre-trends assumptions, for self-fluidity (Left Panel) and market share changes (Right Panel). Fluidity measures the competitive threat from rivals' product market innovations, where higher values indicate greater competitive pressure from potential market entrants. Market share change is computed as growth in firm sales relative to growth in industry total sales, capturing firms' relative market power. A firm-year is classified as high-complementarity if its mean share of data-security specialists participating in non-data-security patent teams over the preceding five years exceeds the 75th percentile of the distribution; this classification is sticky in subsequent years.

ket share is measured as growth in the firm's sales relative to growth in the firm's industry's total sales. This measure increases by up to 200% for high data-complementarity firms by five years after DBNL passage. DBN laws simultaneously reduce market fluidity for the high data-complementarity firms, as their improved operational positions create barriers to entry, limiting the product threat from new competitors in the market. These results indicate the enhanced ability of high data-complementarity firms to leverage their technology advantages to gain market power. The decline in fluidity for high data-complementarity firms post-DBNL enactment underscores the increased barriers to entry, particularly for smaller and less technologically sophisticated competitors. Looking at low data-complementarity firms, we do not see the same pattenrs: while their market shares do (insignificantly) increase in some years, the coefficients are more than twice smaller than for high data-complementarity firms. Furthermore, low data-complementarity firms experience *increases* in fluidity, suggesting increased competitor threats to their product markets. The juxtaposition of the effects for high data-complementarity firms and low data-complementarity firms underscores the importance of expertise cross-sharing for drawing potential competitive advantages from increased salience of data security.

Our findings align with prior theoretical research demonstrating the concentration-

enhancing effects of the data-feedback loop (Farboodi et al. (2019), Farboodi and Veldkamp (2021)). In addition, our results complement the literature highlighting how the integration of artificial intelligence into firm-specific processes enhances both innovation and differentiation (Rock, 2021). Together, these mechanisms illustrate the dual-edged nature of DBNLs: while protecting data integrity, they create disproportionate opportunities for high data-complementarity firms at the expense of industry dynamism.

4.3 Mechanism: Inventor Network and In-House Advantage

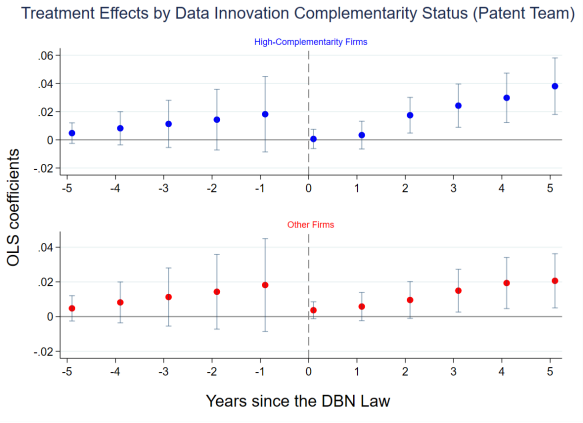
Common engineers and inventors. By construction, the share of data security engineers and inventors in non-data security patent teams is consistently larger (by a factor of 8–10) in high data-complementarity firms relative to low data-complementarity firms. This reflects a significant cross-pollination of expertise and innovation between data security and other technological domains within firms that have high data complementarity. As a result, high data-complementarity firms not only prioritize data security to protect their data-rich environments, but also leverage the specialized knowledge of data security professionals to enhance innovation across different areas of their business. It could imply that these firms recognize the strategic value of integrating data security insights into broader product development and innovation processes, leading to more robust and secure technological solutions.

This interdisciplinary collaboration likely fosters a culture of innovation that is attuned to the complexities of the digital age, where security and functionality are increasingly intertwined. It also points to the potential for high data-complementarity firms to drive industry standards and practices in data security, setting benchmarks that could influence the wider market.

Utilizing existing capabilities or developing new ones? Post DBNLs, firms might leverage existing capabilities to innovate in data security and non-data security technologies. However, the complexity and specificity of the regulations might necessitate new capabilities, leading to an increased share of common inventors to meet advanced data security demands. To better understand this aspect, we examine how the share of data-security innovators in non data-security patent teams evolves after the passage of DBNLs.

Figure 7 suggests that both high and low data-complementarity firms increase the share of data-security inventors in patent teams post-DBNLs. However, the effect is twice as pronounced for high data-complementarity firms (Top panels) as for low data-complementarity firms (Bottom panel). Furthermore, our previous results demonstrate that while both types of firms see an increase in common inventors across patent teams,

Figure 7: Share of Data-Security Innovators in Patent Teams



TThis figure plots BJS dynamic heterogeneous treatment effects of data-security innovator share in patent teams around the staggered adoption of DBN laws ('0' event) across U.S. states. Effects are estimated separately for high data-complementarity firms (Top) and low data-complementarity firms (Bottom), under heterogeneous pre-trends assumptions. Data-security innovators are identified based on our list of inventors with at least 10% of their patent portfolio in data-security patents, where data-security patents are classified using USPTO's CPC codes (detailed in text). A firm-year is classified as high-complementarity if its mean share of data-security specialists participating in non-data-security patent teams over the preceding five years exceeds the 75th percentile of the distribution; this classification is sticky in subsequent years.

only high data-complementarity firms benefit from the cross-pollination. This disparity likely arises because high data-complementarity firms are better positioned to integrate security-focused innovations into their broader data-driven strategies, having already done so to a large extent in the past. This amplifies high data-complementarity firms' productivity and competitive edge, whereas low data-complementarity firms lack pre-existing infrastructure or synergy to fully capitalize on such integration.

More knowledge transfer. High data-complementarity firms likely excel in transferring knowledge between data security and other domains due to their integrated human capital. This integration fosters a cross-pollination of ideas, where advancements in data security can directly influence and improve broader business technologies and vice versa. For example, breakthroughs in encryption can lead to the development of more secure machine learning models.[3] Additionally, the push for explainable and ethical AI demands that data security is embedded within AI algorithms from the ground up, further blurring the lines between specialized innovations. This dynamic environ-

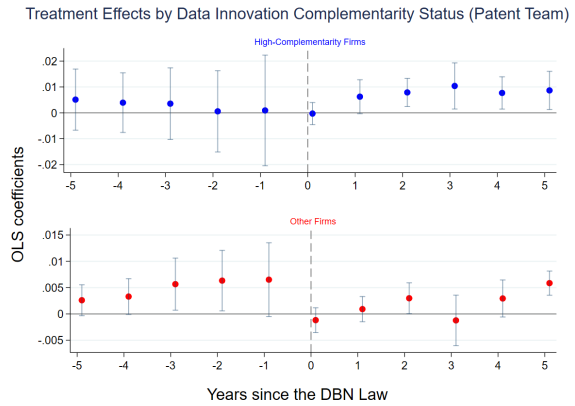
[3]Breakthroughs in encryption can lead to the development of more secure machine learning models by enabling techniques like homomorphic encryption, secure multi-party computation, and differential privacy, which allow computations to be performed on encrypted data without exposing sensitive information. These advances enhance the security and privacy of machine learning systems, particularly in applications involving sensitive data such as healthcare, finance, and personal information, while still allowing robust model training and predictions.

ment creates a synergy where knowledge transfer is not just beneficial but essential for the advancement of both fields.

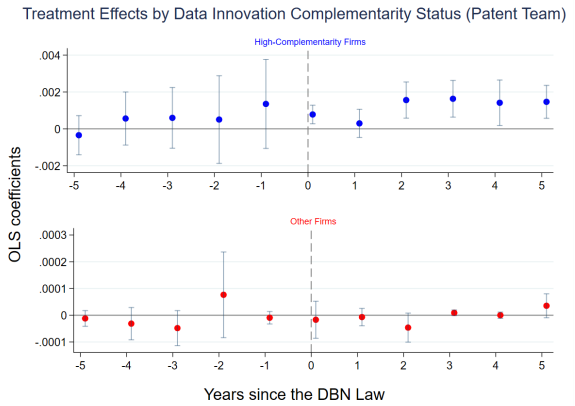
We now test whether high data-complementarity firms exhibit more knowledge transfer between data security and other domains. As shown in Figure 8, high data-complementarity firms (examined in the Top panels) not only increase the share of data-security related patents they cite in their other patents, but also increase the share of *self*-data-security patent citations in their own non-data security patents. This is not the case for low data-complementarity firms, which experience null results. This result suggests that there is a strong knowledge transfer from data protection operations to product and service development in high data-complementarity firms—but not in low data-complementarity firms.

Figure 8: Share of Backward Citations to Data-Security Patents in Non-Data-Security Patents

Left: Share of Data-Security Citations



Right: Share of Data-Security Self-Citations



This figure plots BJS dynamic heterogeneous treatment effects around the staggered adoption of DBN laws ('0' event) across U.S. states. Effects are estimated separately for high data-complementarity firms (Top) and low data-complementarity firms (Bottom), under heterogeneous pre-trends assumptions, for the share of data-security patents in backward citations (Left Panel) and the share of firm's own prior data-security patents in backward citations (Right Panel). Data-security patents are identified using USPTO's CPC codes (detailed in text). A firm-year is classified as high-complementarity if its mean share of data-security specialists participating in non-data-security patent teams over the preceding five years exceeds the 75th percentile of the distribution; this classification is sticky in subsequent years.

Economically, this finding implies that high data-complementarity firms derive significant strategic and competitive advantages from their ability to integrate advancements in data security into broader operations and innovation processes. The increased self-citation of data-security patents in non-data-security patents highlights the importance of data-security-related innovation for the positive effects we observed in the previous subsections. High data-complementarity firms use their expertise in data protection to comply with regulations, and in doing so they create innovative technology that they can leverage as a critical input for developing *other* innovation at the

firm, such as new products and services. This knowledge transfer enhances high data-complementarity firms' productivity frontier and reinforces their position in markets where data-driven innovation is key.

In contrast, low data-complementarity firms lack the infrastructure, complementarities, or absorptive capacity to capitalize on such integration. This disparity may widen the innovation gap and market power concentration, as high data-complementarity firms can further entrench their dominance by making data protection a value-adding component of their offerings, rather than just a compliance necessity. Over time, this dynamic could contribute to reduced market fluidity, higher barriers to entry, and an increasingly polarized economic landscape where leading firms reinforce their advantages through superior knowledge spillovers.

5 The Data Feedback Loop: Modeling Big Data and Data Security Dynamics

In the previous empirical section, we demonstrated that high data-complementarity firms experience an increase in profitability and patenting activity in response to heightened data risk management. This phenomenon is primarily driven by the fact that data risk management compels these firms, which already possess the necessary data-engineering expertise and technological infrastructure, to innovate in data security. These innovations subsequently enhance overall productivity and stimulate further innovation across other areas.

To further understand and rationalize these empirical findings, we now develop a simple growth model of the data economy. This model considers data as a by-product of economic activity that is susceptible to loss due to data risk, and includes firm investments in data security. A critical assumption of the model is that firms differ in the extent to which data security enhances the potential quality of their produced goods. This theoretical framework provides a robust explanation for the observed empirical phenomena. Additionally, we use the model to simulate scenarios that mirror the empirical analysis and perform comparative statics and counterfactual analyses to predict outcomes under different levels of data risk management. We also discuss the policy implications derived from both the theoretical and empirical analyses in a cohesive manner, highlighting how the integration of the two approaches can inform better policy decisions regarding data risk management and innovation promotion.

5.1 Efficient Data Use and Data Security Management

We consider a competitive industry. Time is discrete and infinite. There is a continuum of firms indexed by i. Each firm i produces a good of quality $A_{i,t}$.

$$y_{i,t} = A_{i,t}. \tag{2}$$

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.

Quality $A_{i,t}$ depends on a firm's choice of a production technique $a_{i,t}$, which can be interpreted as managing inventories, or learning about consumer tastes. In each period, and for each firm, there is one optimal technique with a persistent and a transitory component: $\theta_{i,t} + \epsilon_{a,i,t}$. The persistent component $\theta_{i,t}$ is unknown and follows an AR(1) process, where $\eta_{i,t}$ is $i.i.d.$ across time and firms:

$$\theta_{i,t} = \bar{\theta} + \rho(\theta_{i,t-1} - \bar{\theta}) + \eta_{i,t}. \tag{3}$$

Firms have a noisy prior about the realization of θ_0. The transitory shock $\epsilon_{a,i,t}$ is $i.i.d.$ across time and firms and is unlearnable. Deviating from that optimum incurs a quadratic loss in quality:

$$A_{i,t} = \bar{A}_i - (a_{i,t} - \theta_{i,t} - \epsilon_{a,i,t})^2. \tag{4}$$

Quality $A_{i,t}$ is a strictly decreasing function of the difference between the firm's chosen production technique, $a_{i,t}$, and the optimal technique $\theta_{i,t} + \epsilon_{a,i,t}$. A decreasing function means that techniques far away from the optimum result in inferior quality goods.

Data as by-product. Data helps firms infer $\theta_{i,t}$. The term ϵ_a indicates that firms are incapable of fully inferring $\theta_{i,t}$ at the end of each period, making the accumulation of past data a valuable asset. If a firm knew the current value of $\theta_{i,t}$, it would maximize quality by setting $a_{i,t} = \theta_{i,t}$.

In our model, similar to Farboodi et al. (2019) and Farboodi and Veldkamp (2021), data is a by-product of economic activity. Each firm passively obtains z data points as a by-product of production. Each data point $m \in [1 : z]$ reveals

$$s_{i,t,m} = \theta_{i,t} + \epsilon_{i,t,m}, \tag{5}$$

where $\epsilon_{i,t,m}$ is $i.i.d.$ across firms, time, and signals. For tractability, we assume that all the shocks are normally distributed: fundamental uncertainty is $\eta_{i,t} \sim N(\mu, \sigma_\theta^2)$, signal

noise is $\epsilon_{i,t,m} \sim N(0, \sigma_\epsilon^2)$, and the unlearnable quality shock is $\epsilon_{a,i,t} \sim N(0, \sigma_a^2)$.

Data risk. Data is subject to data risk or data incident risk, meaning that it can be lost and, in that case, it can no longer be used for prediction. We denote the degree of data risk by $\vartheta \in [0, 1]$. With probability ϑ, a firm risks losing all its data, while with probability $(1 - \vartheta)$ the firm keeps its data generated as a by-product of activity, $z\sigma_\theta^2$. Thus, the data endowment under data risk is $(1 - \vartheta)z\sigma_\epsilon^2$.

Data security. A key assumption of our model is that firms are heterogeneous in their capability to protect themselves against data risk. High-data-complementarity (H-type) firms can develop in-house data security protection, while low complementarity (L-type) firms cannot develop this security internally, but can buy it externally from high-data-complementarity (H-type) firms.

The essential distinction between in-house and external data security is that internal data security can also be used to innovate, apart from providing protection against data loss and destruction. This is because in-house data security is typically more easily integrated with existing R&D and product development systems, and tends to be more tailored for a firm's specific business needs. In the model, innovation is modeled as an increase in the productivity ceiling $\overline{A}_i$.

Low-complementarity firms do not generate in-house security, but they can buy it externally from high-data-complementarity (H-type) firms. In this case, they can only use it to mitigate the impact of data risk and not to innovate (i.e., they can use the security software for protecting their production process, but their R&D department does not know and is unable to use the security software for product improvements).

Let m_H represent the share of high-data-complementarity (H-type) firms. Aggregate output is then the sum of weighted outputs for the two types of firms:

$$Y_t = \int_0^1 A_{i,t} di = m_H A_{H,t} + (1 - m_H) A_{L,t}. \tag{6}$$

Let $\tau_t \geq 0$ represent the investment in in-house data security made by a high-data-complementarity (H-type) firm. Let also $\delta_t \geq 0$ represent the amount of external data security bought by a L-type firm from the high-data-complementarity (H-type) firm at an endogenous price denoted by π. Given the firm shares, the amount of protection that is sold by a high-data-complementarity (H-type) producer must be $\frac{1-m_H}{m_H}\delta_t$. In this case, on the aggregate H-type firms sell $(1 - m_H)\delta_t$, which is precisely the value of protection purchased by L-type firms.

Non-rivlary. When a company invests in data security measures such as firewalls,

encryption protocols, or security software, these measures protect the company's data and systems without necessarily reducing their effectiveness for other companies that may use similar security tools. This suggests that data security is (partially) non-rival.

Thus, we assume that when a high-data-complementarity (H-type) firm sells a given amount of data protection, it retains, for its own use, a share $1-\iota$ of such protection, where $\iota \in (0,1)$. Therefore, the H-firm that invests τ_t in data security and trades $\frac{1-m_H}{m_H}\delta_t \leq \tau_t$, will retain, for its own use, $\tau_t - \iota\frac{1-m_H}{m_H}\delta_t$. This amount of data protection can be used to mitigate the impact of data risk, transforming the term $(1-\vartheta)z$ into $\left[1-\vartheta e^{-\left(\tau_t-\iota\frac{1-m_H}{m_H}\delta_t\right)}\right]z$. Note that if $\tau_t - \iota\frac{1-m_H}{m_H}\delta_t = 0$, there is no use of data protection, and the effect of data risk over data is maximum; if $\tau_t - \iota\frac{1-m_H}{m_H}\delta_t \to \infty$, then there is full protection, and the original data endowment maintains its integrity.

Firm problem. With this in mind, we can write firm i's optimization problem, where $i \in \{H, L\}$. As mentioned previously, the high-data-complementarity (H-type) firm can use the investment in data security to enhance the potential quality of the produced good. Hence, constant $\overline{A}_i$ is replaced, for this type of firm, by the term $\overline{A}e^{b\left(\tau_t-\iota\frac{1-m_H}{m_H}\delta_t\right)}$.

A high-data-complementarity (H-type) firm chooses a sequence of quality decisions $a_{i,t}$, in-house data security investments τ_t, and how much data security δ_t to sell at price π_t to maximize:

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\overline{A}e^{b\left(\tau_t-\iota\frac{1-m_H}{m_H}\delta_t\right)} - (a_{i,t} - \theta_{i,t} - \epsilon_{a,i,t})^2 - \tau_t + \frac{1-m_H}{m_H}\delta_t\pi_t - r\right] \tag{7}$$

An L-type firm chooses a sequence of quality decisions $a_{i,t}$, and how much external data protection δ_t to buy at price π_t to maximize:

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left[\bar{A} - (a_{i,t} - \theta_{i,t} - \epsilon_{a,i,t})^2 - \delta_t\pi_t - r\right] \tag{8}$$

Note the differences between the two expressions: innovation from data security is possible for the high-data-complementarity (H-type) firm but not for the L-type firm ; the cost of investment in data security is present only in the H-type firm expression; protection trading is a revenue for those who sell it and a cost for those who buy it.

The stock of knowledge. The information set of firm $i \in \{H, L\}$ when it chooses its technique $a_{i,t}$ is $\mathcal{I}_{i,t} = \{\mathcal{I}_{i,t-1}, \{s_{i,t-1,m}\}_{m=1}^{z}, A_{i,t-1}\}$ where z is the net numbers of points added each period as a by-product of economic activity. To make the problem

recursive, we construct a helpful summary statistic for this information, called the 'stock of knowledge.' A firm's stock of knowledge is the inverse of its posterior variance, or in other words, the precision of firm i's forecast of θ_t, which is formally:

$$\Omega_{i,t} = \mathbb{E}\left[(\mathbb{E}[\theta_t|\mathcal{I}_{i,t}] - \theta_t)^2\right]^{-1} \tag{9}$$

Note that the interior of the expression is the difference between a forecast, $\mathbb{E}[\theta_t|\mathcal{I}_{i,t}]$ and the realized value, θ_t, and is therefore a forecast error. An expected squared forecast error is the variance of the forecast. It is also called the variance of θ_t, conditional on the information set $\mathcal{I}_{i,t}$, or the posterior variance. The inverse of a variance is a precision. Thus, this is the precision of firm i's forecast of θ_t.

A law of motion for knowledge The state variables of the recursive problems in (7) and (8) are the prior mean and variance of beliefs about $\theta_{i,t-1}$, and the new data points. Taking a first order condition with respect to the technique choice, we find that the optimal technique is $a^*_{i,t} = \mathbb{E}_i[\theta_{i,t}|\mathcal{I}_{i,t}]$. Given the posterior variance of beliefs in (9), the expected quality for the high-data-complementarity (H-type) and the L-type firms, respectively, are

$$\mathbb{E}[A_{H,t}] = \overline{A}e^{b\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)} - \Omega^{-1}_{H,t} - \sigma^2_a \tag{10}$$

$$\mathbb{E}[A_{L,t}] = \overline{A} - \Omega^{-1}_{L,t} - \sigma^2_a \tag{11}$$

Deriving the law of motion for the stock of knowledge, $\Omega_{i,t}$, requires adding new data from two sources: 1) data as a by-product of production, which is subject to data risk but can be protected through data security and 2) data inferred from a firm observing its own quality at the end of a production period. These two pieces of information are incorporated into beliefs using Bayes' law.

Each firm $i \in \{H, L\}$ observes $z_i = z$ data points as a by-product of economic activity. This means that the sum of the precisions of all the signals (data points), $z_i\sigma_\epsilon^{-2}$ is part of the stock of knowledge. Both types of firms, the H-type and the L-type, are subject to data risk, which can be reduced through protection. The high-data-complementarity (H-type) firm reduces data risk by the amount of data security it retains for its own use, $\tau_t - \iota\frac{1-m_H}{m_H}\delta_t \leq \tau_t$, after it invests τ_t in data security and trades $\frac{1-m_H}{m_H}\delta_t \leq \tau_t$ data protection which is non-rival. This amount of data protection can be used to mitigate the impact of data risk, implying that the weighted sum of precisions of data points obtained as a byproduct of economic activity, subject to data risk and after optimal data security decisions, is $\left[1 - \vartheta e^{-\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)}\right] z\sigma_\epsilon^{-2}$. The L-type firm buys protection in amount δ_t and, therefore, the weighted sum of precisions of data points

obtained as a byproduct of economic activity, subject to data risk and after optimal data security decisions, is $\left[1-\vartheta e^{-\delta_t}\right] z\sigma_\epsilon^{-2}$.

Moreover, each firm $i \in \{H, L\}$ also learns from seeing its own realization of quality $A_{i,t}$ at the end of each period t, with precision σ_a^{-2}. This information is different from the produced data because the quality realization is a signal about θ_t, not about θ_{t+1}. Therefore, σ_a^{-2} gets added to the time-t stock of knowledge and depreciates, just like other time-t knowledge that the firm takes with it to time $t+1$.

Lemma 1 expresses the dynamic knowledge constraint that puts together data depreciation and data inflows.

Lemma 1 *The dynamic knowledge constraint is, for the high-data-complementarity (H-type) firm:*

$$\Omega_{H,t+1} = \left[\rho^2(\Omega_{H,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + \left[1 - \vartheta e^{-\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)}\right] z\sigma_\epsilon^{-2} \tag{12}$$

The L-type firm buys protection in amount δ_t and, therefore,

$$\Omega_{L,t+1} = \left[\rho^2(\Omega_{L,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + \left(1 - \vartheta e^{-\delta_t}\right) z\sigma_\epsilon^{-2} \tag{13}$$

In this last case, if the firm buys no protection, data loss risk occurs in a share ϑ; if it buys infinite protection, it faces no data risk.

The demonstration for this lemma and all subsequent lemmas and propositions can be found in Appendix C. The proof involves utilizing Bayes' law, or alternatively, the Ricatti equation within a modified Kalman filter framework. Given the similarity in information structure to that of a Kalman filter, the sequence of conditional variances (or conversely, their inverses, the sequence of precisions) is deterministic.

5.2 Recursive Problem, Equilibrium and Steady State

Lemma 2 proceeds with the recursive representation of the expected firm value.

Lemma 2 *The optimal sequences of in-house data security investments $\{\tau_t\}$ and data security sales $\{\delta_t\}$ solve the following current-value Hamiltonian function for the high-data-complementarity (H-type) firm:*

$$H_{H,t}(\Omega_{H,t}, \tau_t, \delta_t, p_{H,t}) = \overline{A} e^{b\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)} - \Omega_{H,t}^{-1} - \sigma_a^2 - \tau_t + \frac{1-m_H}{m_H}\delta_t\pi_t - r + \tag{14}$$
$$+ \beta p_{H,t+1}(\Omega_{H,t+1} - \Omega_{H,t})$$

$$\textit{where} \quad \Omega_{H,t+1} = \left[\rho^2(\Omega_{H,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + \left[1 - \vartheta e^{-\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)}\right] z\sigma_\epsilon^{-2} \tag{15}$$

and $p_{H,t}$ is the shadow-price or co-state variable associated with the state variable, and the transversality condition is $\lim_{t\to\infty} \Omega_{H,t}\beta^t p_{H,t} = 0$.

The optimal sequence of data security purchases $\{\delta_t\}$ solves the following current-value Hamiltonian function for the L-type firm:

$$H_{L,t}(\Omega_{L,t}, \tau_t, \delta_t, p_{L,t}) = \overline{A} - \Omega_{L,t}^{-1} - \sigma_a^2 - \delta_t \pi_t - r + \beta p_{L,t+1}(\Omega_{L,t+1} - \Omega_{L,t}) \tag{16}$$

$$\text{where} \quad \Omega_{L,t+1} = \left[\rho^2(\Omega_{L,t} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + \left(1 - \vartheta e^{-\delta_t}\right) z\sigma_\epsilon^{-2} \tag{17}$$

and $p_{L,t}$ is the shadow-price or co-state variable associated with the state variable, and the transversality condition is $\lim_{t\to\infty} \Omega_{L,t}\beta^t p_{L,t} = 0$.

See the Appendix for the proof. This result greatly simplifies the problem by collapsing it to a deterministic dynamic system involving only one state variable, $\Omega_{i,t}$, where $i = H$ or $i = L$. The reason we can do this is that quality $A_{i,t}$ depends on the conditional variance of $\theta_{i,t}$ and because the information structure is similar to that of a Kalman filter, where the sequence of conditional variances is generally deterministic.[4] This Kalman system has a 2-by-1 observation equation, with $n_{i,t} = z$ signals about $\theta_{i,t}$ and one signal about $\theta_{i,t-1}$. The signal about $\theta_{i,t-1}$ comes from observing last period's output, which reveals quality $A_{i,t-1}$, which, in turn, reveals $\theta_{i,t} + \epsilon_{a,i,t}$.[5]

Equilibrium. From the Hamiltonian functions, and assuming all variances are equal such that $\sigma_\theta^2 = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2$, we can derive the equilibrium conditions.

$$\frac{\partial H_{H,t}}{\partial \tau_t} = 0 \Rightarrow \beta p_{H,t+1} = \frac{1 - b\overline{A}e^{b\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)}}{\vartheta e^{-\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)} z\sigma^{-2}} \tag{18}$$

$$\frac{\partial H}{\partial \delta_t} = 0 \Rightarrow \beta p_{H,t+1} = \frac{\pi_t - b\overline{A}\iota e^{b\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)}}{\vartheta\iota e^{-\left(\tau_t - \iota\frac{1-m_H}{m_H}\delta_t\right)} z\sigma^{-2}} \tag{19}$$

$$\beta p_{H,t+1} - p_{H,t} = -\frac{\partial H}{\partial \Omega_{H,t}} \Rightarrow \left[\rho + \frac{\sigma^2}{\rho}(\Omega_{H,t} + \sigma^{-2})\right]^{-2} \beta p_{H,t+1} = p_{H,t} - \Omega_{H,t}^{-2} \tag{20}$$

From (31) and (32), a constant optimal trading price, which is simply $\pi_t = \iota$,

[4]The optimal choice of technique is always the same: $a_{i,t}^* = \mathbb{E}_i[\theta_{i,t}|\mathcal{I}_{i,t}]$. The way $a_{i,t}$ enters into expected quality $A_{i,t}$ is through $\mathbb{E}[(\mathbb{E}[\theta_{i,t}|\mathcal{I}_{i,t}] - \theta_{i,t})^2]$, which is the conditional variance $\Omega_{i,t}$. We can replace the entire sequence of $a_{i,t}^*$ with the sequence of variances, which is deterministic here because of normality. The only randomness in this model comes from the signals and their realizations, but they never affect the conditional variance, since normal means and variances are independent. Thus, given $\Omega_{i,t-1}$, $\Omega_{i,t}$ is a sufficient statistic for $n_{i,t} = z$ and $\Omega_{i,t+1}$. The mean $\mathbb{E}[\theta_{i,t}|\mathcal{I}_{i,t}]$ is not a state variable because it only matters for determining $a_{i,t}$ and does not affect anything else.

[5]Firms observe $(\theta_{i,t} + \epsilon_{a,i,t})^2$. For tractability, we assume that firms know whether the root is positive or negative. For the derivation of the belief updating equations, see the Appendix.

emerges. The price of protection is directly associated with the degree of its own non-rivalry. If protection is completely non-rival (i.e., $\iota = 0$), then its price is zero; if protection is fully rival, its price is 1.

For the L-type firm, the equilibrium conditions are:

$$\frac{\partial H}{\partial \delta_t} = 0 \Rightarrow \beta p_{L,t+1} = \frac{\pi_t}{\vartheta e^{-\delta_t} z \sigma^{-2}} \tag{21}$$

$$\beta p_{L,t+1} - p_{L,t} = -\frac{\partial H}{\partial \Omega_{L,t}} \Rightarrow \left[\rho + \frac{\sigma^2}{\rho}(\Omega_{L,t} + \sigma^{-2})\right]^{-2} \beta p_{L,t+1} = p_{L,t} - \Omega_{L,t}^{-2} \tag{22}$$

Steady-state. The steady-state of the economy is characterized by a level of data security held by high-data-complementarity (H-type) firms after trade given by:

$$\tau^* - \iota \frac{1 - m_H}{m_H} \delta^* = -\ln\left(\frac{z - \Xi_H}{\vartheta z}\right) \tag{23}$$

where $\Xi_H \equiv \left\{\Omega_H^* - [\rho^2(\Omega_H^* + \sigma^{-2})^{-1} + \sigma^2]^{-1}\right\} \sigma^2$. At steady-state, the amount of protection bought by L-type firms is given by:

$$\delta^* = -\ln\left(\frac{z - \Xi_L}{\vartheta z}\right) \tag{24}$$

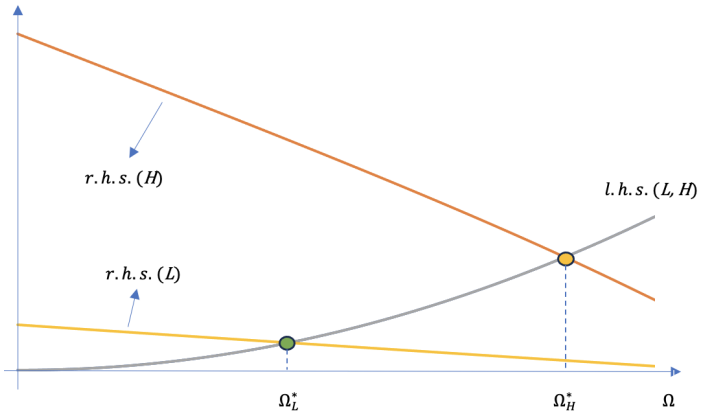
with $\Xi_L \equiv \left\{\Omega_L^* - [\rho^2(\Omega_L^* + \sigma^{-2})^{-1} + \sigma^2]^{-1}\right\} \sigma^2$.

Figure (9) plots the equilibrium knowledge levels of this economy. The demand and supply of knowledge for high-data-complementarity (H-type) firms intersect at a higher level than the demand and supply of knowledge for L-type firms. The demand of L-type firms is flatter and more inelastic than the demand of H-type firms. Thus, in equilibrium, high-data-complementarity (H-type) firms end up with a higher level of knowledge than L-type firms.

Table 2 illustrates the steady-state equilibrium of this economy. In this equilibrium, high-data-complementarity (H-type) firms invest 1.296 in in-house data protection, sell 0.130 worth of data protection to low-data-complementarity (L-type) firms, and maintain a data protection level of 0.335, which is significantly higher than the L-type firms' protection level of 0.130. This higher investment in data protection by H-type firms aligns with our empirical observation that high data-complementarity firms are better equipped to handle data risk, translating their data protection efforts into enhanced innovation and productivity.

Furthermore, our model shows that in steady-state, knowledge, quality, and profits are all higher for H-type firms compared to L-type firms. This theoretical outcome is

Figure 9: Steady-state stocks of knowledge



Legend: The figure shows the equilibrium levels of knowledge for high-data-complementarity (H-type) firms (in orange on the right) and L-type firms (in green on the left) as a function of the data risk index, ϑ, on the X-axis. High-data-complementarity (H-type) firms achieve a higher level of steady-state knowledge than L-type firms. The parameters used in this simulation are the following: $z = 10$, $\rho = 0.9$, $\sigma_\theta^2 = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2 = 2.5$, $m_H = 1/3$, $\iota = 0.6$, $\beta = 0.96$, $\vartheta = 0.75$, $\overline{A} = 25$, $b = 0.035$, and $r = 1$.

consistent with our empirical results, which demonstrate that high data-complementarity firms not only mitigate the adverse impacts of data risk but also harness these challenges to amplify their innovation activities. This dual benefit of data protection—preserving data and driving innovation—highlights the strategic advantage that high data complementarity firms possess, as evidenced by their superior performance metrics in our empirical analysis.

Overall, the theoretical model supports and rationalizes the empirical findings, providing a robust framework that explains how firms with high data complementarity can turn data risks into opportunities for growth and competitive advantage. This synergy between our empirical and theoretical analyses underscores the importance of strategic investments in data security for fostering long-term innovation and profitability.

5.3 Results and Implications

Throughout this section, we use a numerical example to highlight some model implications. These results comprise the impact of data salience on firm profits, aggregate output, and on the timing of the decision to engage in data security protection.

5.3.1 Data Protection Helps Firms Hedge Data Risk

Our first numerical experiment studies how an increase in firm data risk changes firm profitability. We start by simulating firm profits in a model with no data protection. Then, we turn on data security protection for both types of firms to observe how their

Table 2: Steady-state

Parameter	**Symbol**	**Steady-state**
Knowledge H-type	$\Omega_H^\star$	3.224
Knowledge L-type	$\Omega_L^\star$	1.609
In-house data protection	$\tau^\star$	1.296
Data security traded	$\delta^\star$	0.130
Quality H-type	A_H^*	23.207
Quality L-type	A_L^*	21.879
Profits H-type	Π_H^*	21.068
Profits L-type	Π_L^*	20.800
Total output	Y	22.321

Legend: The parameters used in this simulation are the following: $z = 10$, $\rho = 0.9$, $\sigma_\theta^2 = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2 = 2.5$, $m_H = 1/3$, $\iota = 0.6$, $\beta = 0.96$, $\vartheta = 0.75$, $\overline{A} = 25$, $b = 0.035$, and $r = 1$.

profits change. To compute the change in firm profitability when firms face increasingly higher data risk, we change the data risk index ϑ continuously from no data risk ($\vartheta = 0$) to maximum data risk ($\vartheta = 1$) and re-compute the steady state. Figure 10 shows that the profits of H-type firms with data security fall by less than the profits of L-type firms as data risk increases. Moreover, the profits of both types of firms with no data security protection at all drop dramatically as the overall level of data risk increases in the economy.

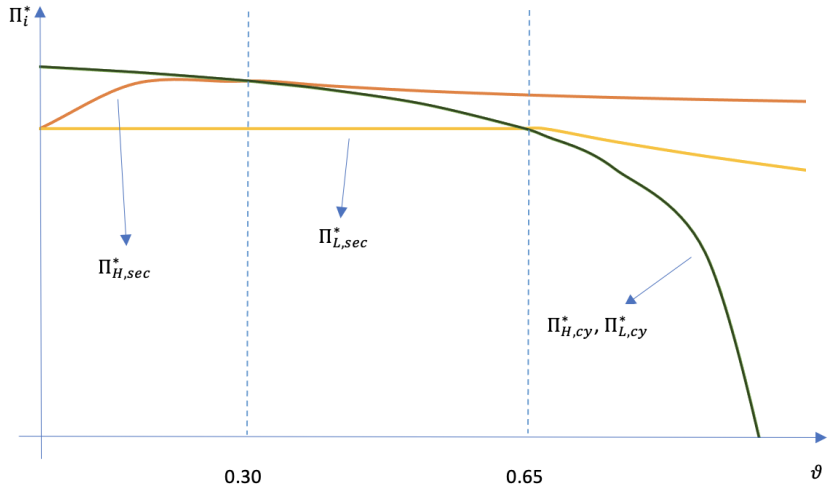
Without protection, the profits (in green) of H-type firms are the same as of L-type firms and they decrease in the data risk index ϑ. Initially, the profits without data security decline slowly, but after the second threshold, they decline rapidly because the cost incurred in knowledge loss increases exponentially with data risk without protection. With protection, however, the profits of H-type firms (in orange) are always higher than of L-type firms (in yellow). As data risk increases, the profits of H-type firms decrease at a smaller rate than of L-type firms (in yellow). An interesting observation is that initially, with protection, the profits of H-type firms first increase because the benefit of protection (which is a data security-driven innovation) is initially higher than the cost of cyber crime.

5.3.2 High Complementarity Firms Engage in Protection at Lower Risk Levels Than Low Complementarity Firms

What governs the steady-state size of firms is firm data security levels as a function of the data risk index, ϑ, plotted in Figure 11.

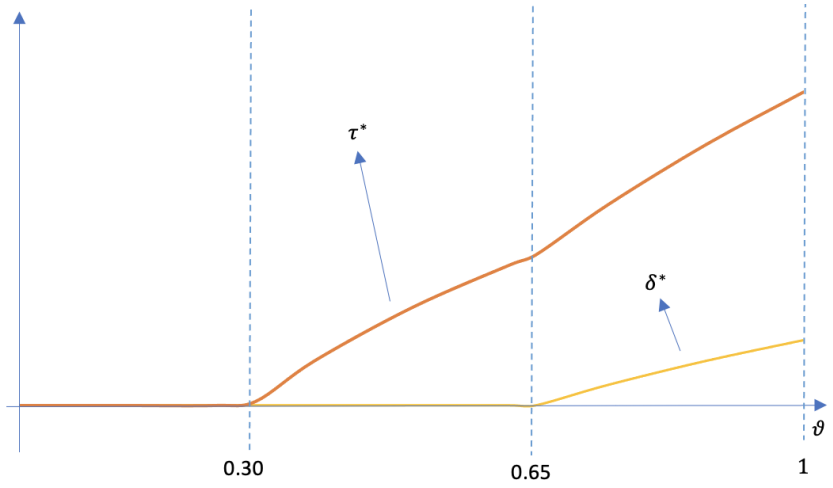
Evaluating the model for different values of ϑ, and letting all other parameters be as before, we find two critical thresholds: at $\vartheta = 0.6583$, optimal data security purchases,

Figure 10: Profits as a function of data risk



Legend: This figure plots the steady-state profit levels for H-type firms with (in orange, $\Pi^{\star}_{H,sec}$) and without data protection (in green, $\Pi^{\star}_{H,cy}$), and L-type firms with (in yellow, $\Pi^{\star}_{L,sec}$) and without data protection (in green, $\Pi^{\star}_{L,cy}$), as a function of the data risk index, ϑ, on the X-axis. The parameters used in this simulation are the following: the data endowment $z = 10$, the coefficient of the AR(1) process $\rho = 0.9$, all variances $\sigma^2_\theta = \sigma^2_a = \sigma^2_\epsilon = \sigma^2 = 2.5$, the share of H-type firms $m_H = 1/3$, the non-rivalry parameter $\iota = 0.6$, the inter-temporal discount factor $\beta = 0.96$, the maximum quality threshold $\overline{A} = 25$, the innovation externality $b = 0.035$, and the cost of capital $r = 1$.

Figure 11: Data security as a function of data risk



Legend: The figure plots in-house data security investment, τ_t, by H-type firms (in orange), and external data security acquisition by L-type firms (in yellow). Notice the two critical thresholds at which in-house data security and external data security become strictly positive. The parameters used in this simulation are the following: the data endowment $z = 10$, the coefficient of the AR(1) process $\rho = 0.9$, all variances $\sigma^2_\theta = \sigma^2_a = \sigma^2_\epsilon = \sigma^2 = 2.5$, the share of H-type firms $m_H = 1/3$, the non-rivalry parameter $\iota = 0.6$, the inter-temporal discount factor $\beta = 0.96$, the maximum quality threshold $\overline{A} = 25$, the innovation externality $b = 0.035$, and the cost of capital $r = 1$.

δ^*, changes from negative to positive, implying that L-type firms buy protection only for $\vartheta > 0.6583$. For $\vartheta \leq 0.6583$, H-type firms have to choose whether to invest in

protection, knowing that they cannot sell data protection. H-type firms are indifferent between investing in protection or not at a critical threshold level of $\vartheta = 0.3$. For $\vartheta > 0.3$, H-type firms invest in protection, otherwise they do not.

Focusing on the middle-interval in which high data-complementarity (H-type) firms invest in data security and low data-complementarity (L-type) firms do not, a back-of-the-envelope calculation using our non-calibrated parameters generates a 45% increase in data security investments for H-type firms and a 0% increase in data security for L-type firms for a standard deviation increase in data risk.[6] This result should be interpreted qualitatively, not quantitatively, because our simple exercise is to use comparative statics exercises to provide intuition, not to carefully calibrate and take the model to the data. However, this qualitative theoretical result is consistent with the empirical evidence from Subsection 4.1, which demonstrates that high data-complementarity firms increase their citation-weighted patent counts while low data-complementarity firms do not as a reaction to heightened data risk, proxied by DBNLs.

5.3.3 Data Complementarity Can Also Sustain Growth

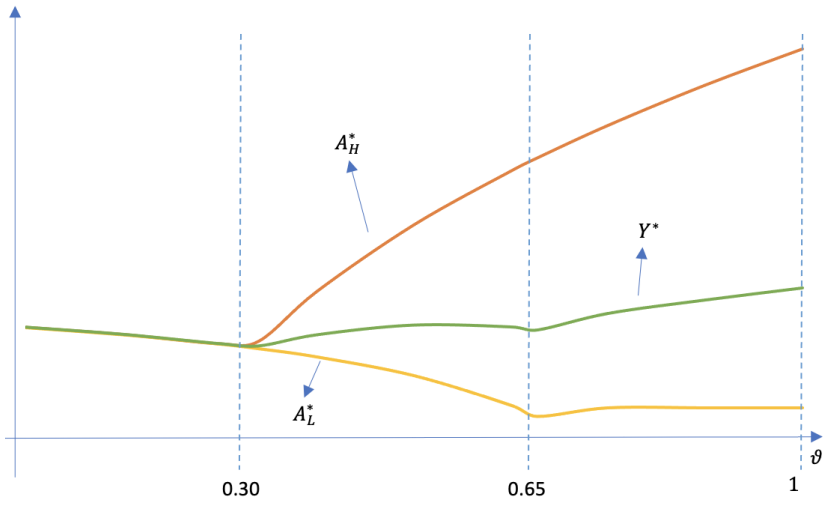
Surprisingly, while one expects aggregate economic output to decrease in data risk, there is a counteracting force that works especially at high levels of risk. This is shown in Figure 12.

Firms with a high capacity for in-house data security protection (in orange) use this protection to innovate, which raises the quality and quantity of production. Output increases in data risk for H-type firms at moderate to high levels of data risk. L-type firms do not have this positive spillover because they only use data security for their own protection to mitigate the negative effects of data risk. The aggregate output is a weighted average of the output of the two types of firms. Concerning the evolution of Y^* as ϑ increases, one notices that an initial fall is counteracted when H-type firms start to invest in protection, and this process gains momentum when L-type firms start to protect their data as well.

Focusing on the middle interval where high data-complementarity (H-type) firms invest in data security and grow, while low data-complementarity (L-type) firms do not invest and contract rapidly, we observe a significant divergence in output. Using non-calibrated parameters for a back-of-the-envelope calculation, we find a 41% increase in output for H-type firms and a -14% decrease in output for L-type firms for a standard

[6] The slope of the data security investment is 2.71 for H-type firms, and 0 for L-type firms. Assuming that both H-types and L-types are normally distributed on the unit interval ϑ, this translates into an increase of 45% [$= 1/6 * slope = 1/6 * 2.71$] for H-types and an increase of 0% for L-type firms.

Figure 12: Output as a function of data risk



Legend: The parameters used in this simulation are the following: the data endowment $z = 10$, the coefficent of the AR(1) process $\rho = 0.9$, all variances $\sigma_\theta^2 = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2 = 2.5$, the share of H-type firms $m_H = 1/3$, the non-rivalry parameter $\iota = 0.6$, the inter-temporal discount factor $\beta = 0.96$, the maximum quality threshold $\overline{A} = 25$, the innovation externality $b = 0.035$, and the cost of capital $r = 1$.

deviation increase in data risk. [7]

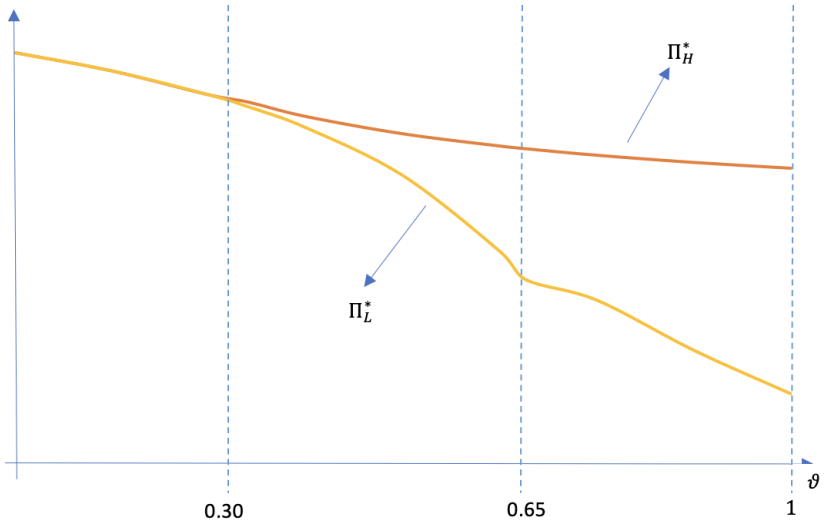
This result should be interpreted qualitatively rather than quantitatively, as our exercise is intended to use comparative statics to provide intuition rather than to precisely calibrate and apply the model to the data. However, this qualitative theoretical result aligns with the empirical evidence showing that high data-complementarity firms increase their sales and profitability in response to heightened data risk, as proxied by DBNLs, whereas low data-complementarity firms do not experience similar gains.

We can also recover the representation of profits in Figure 10, to plot the actual profits, given the choices of firms on whether to obtain data protection. Figure 13 clarifies again the existence of three stages and the fact that data risk is much less harmful for H-type firms because these make use of the innovation externality that data security allows for.

The model is simple, but it generates some powerful predictions. Data risk hurts firms in the modern economy and firms make lower profits at increasingly high levels of risk. However, there is a silver lining: data risk can sustain growth and innovation when it allows firms to use data security for innovation. We allowed some firms in the economy the potential to use data security to improve their productivity ceiling. When given this opportunity, data risk can sustain firm growth and innovation because there

[7] The slope of the output function is 2.46 for H-type firms and -0.82 for L-type firms. Assuming a normal distribution on the unit interval ϑ for both types, this translates into a 41% increase [$= 1/6 \times 2.46$] for H-types and a -14% decrease [$= 1/6 \times (-0.82)$] for L-type firms.

Figure 13: Realized (equilibrium) profits



Legend: The parameters used in this simulations are the following: the data endowment $z = 10$, the coefficient of the AR(1) process $\rho = 0.9$, all variances $\sigma_\theta^2 = \sigma_a^2 = \sigma_\epsilon^2 = \sigma^2 = 2.5$, the share of H-type firms $m_H = 1/3$, the non-rivalry parameter $\iota = 0.6$, the inter-temporal discount factor $\beta = 0.96$, the maximum quality threshold $\overline{A} = 25$, the innovation externality $b = 0.035$, and the cost of capital $r = 1$.

are innovation externalities that arise from data risk protection.

5.4 Discussion and Policy Implications

The empirical findings from Section 4.1 reveal that high data-complementarity firms show a significant increase in innovation and profitability when faced with heightened data risk. These firms, equipped with advanced data-engineering capabilities and technological infrastructure, are able to develop in-house security solutions, turning data risks into opportunities for innovation. This empirical evidence aligns closely with our theoretical model, which predicts that firms investing more in data protection will see greater innovation and growth.

In Section 5.3, we delved into the results and their broader implications and connection to the empirical findings. The steady-state equilibrium, as detailed in Section 5.2, demonstrates that H-type firms (high data-complementarity) invest significantly more in in-house data protection compared to L-type firms (low data-complementarity) on average. This higher level of investment not only safeguards their data but also facilitates innovation spillovers, enhancing the overall quality and profitability of their products.

Our comparative statics results from Section 5.2 examining the steady-state of an economy with increasing higher levels of data risk support our empirical insights as well, showing that high data-complementarity firms experience a notable increase in

innovation and output growth in response to heightened data risk. In the empirical analysis, a one standard deviation increase in data salience correlates with a 7% rise in overall patent counts and a 5.5% increase in non-data-security-related patents. A back-of-the-envelope calculation using the middle-range of data salience from our theoretical model corresponds to a 45% increase in data security investments and a 41% increase in output for high data-complementarity (H-type) firms, and a 0% increase in data security and a -14% decrease in output for low data-complementarity (L-type) firms. These effects change in a high data-salience environment to a 49% increase in data security investments and a 24% increase in output for high data-complementarity (H-type) firms, and a 19% increase in data security and a 0% decrease in output for low data-complementarity (L-type) firms. This result should be viewed qualitatively rather than quantitatively, as our use of comparative statics is meant to provide intuition rather than precise calibration or application to empirical data. Nevertheless, this qualitative theoretical outcome is consistent with the empirical evidence, which shows that high data-complementarity firms increase their sales and profitability in response to heightened data risk, as indicated by DBNLs, while low data-complementarity firms do not exhibit similar improvements.

The synergy between the theoretical predictions and the empirical validation underscores the dual role of data protection in both preserving existing data and driving growth and innovation in the presence of heightened data risk, which is a key prediction of our theoretical model and a key finding of our empirical analysis.

Moreover, the implications of these findings are significant. First, they suggest that differences between high data-complementarity and low data-complementarity firms and industries magnify with increasing data security risks. Second, they suggest that policies aimed at enhancing data security should consider the heterogeneous nature of firms and industries as well as the levels of data risk that different industries are exposed to. Third, for high data-complementarity sectors, encouraging investments in data protection can spur innovation and economic growth. However, for sectors like health and accommodation, additional support may be needed to mitigate the operational vulnerabilities and compliance costs associated with data risk. Fourth, in light of these challenges, the need for innovation, collaboration, diversification of digital defense strategies and resilience to breaches that can lead to data loss and destruction is greater than ever. Raising awareness among less sophisticated, low data-complementarity firms is essential given the potential huge costs that data breaches entail.

One public policy solution could be the active transfer of knowledge from high data-complementarity firms, that have extensive experience in defending against data breaches, with the aim of transferring knowledge and best practices to the world of

SMEs and public authorities, which are typically less high data-complementarity entities. Such collaboration, supported by government incentives such as grants or tax breaks, could significantly increase the cyber resilience of these vulnerable sectors.

Lastly, the private sector, namely the insurance industry, can play a critical role by developing standardized cyber insurance policies that encourage low data-complementarity firms to adopt and maintain high data security standards. Such policies can serve as a catalyst for widespread compliance with robust data protection measures by linking premiums and coverage to the implementation of specific data security protocols. A unified approach using both government support and private market solutions would provide a promising way to mitigate data risks and build a secure, resilient digital infrastructure for AI- and low data-complementarity companies alike.

In conclusion, the integration of our empirical findings with the theoretical model provides a comprehensive understanding of how data risk impacts firm behavior and performance and sheds light on potential policy solutions. This synergy between theory and empirics not only validates our model but also offers actionable insights for policymakers and industry leaders aiming to foster innovation and resilience in the digital economy.

6 Conclusion

This paper highlights the transformative role of data-feedback loops in shaping firm growth, innovation, and market power within the data-driven economy. By leveraging the staggered adoption of Data Breach Notification Laws (DBNLs) as a quasi-exogenous shock, we demonstrate that heightened data risk amplifies the benefits of data complementarity. Our findings reveal that high data-complementarity firms are uniquely positioned to turn regulatory and operational challenges into opportunities for innovation and productivity gains, reinforcing their competitive advantages.

Specifically, high data-complementarity firms respond to increased data salience by intensifying their innovation efforts, both in data security and product development. A key mechanism driving this dynamic is the dual role of data-related inventors, whose expertise in data protection directly feeds into broader innovation activities, exemplifying the reinforcing nature of the data-feedback loop. In contrast, low data-complementarity firms face stagnation or even decline, widening the performance gap between the two groups.

Our study makes three primary contributions. First, it advances the literature on valuation of data and firm performance by formalizing and empirically validating the data-feedback loop as a driver of innovation and growth. To do so, we construct a novel measure of firm-level complementarity between data-related and non-data-related

human capital, which is especially critical in the modern data economy. Second, our paper introduces a new perspective on the interplay between data security and product innovation, highlighting the critical role of interdisciplinary data expertise. Finally, we provide novel evidence on the broader economic and competitive implications of data protection regulations, emphasizing their dual role as both enablers of technological advancement and potential barriers to industry dynamism.

Future research could explore how these dynamics evolve as big data continues to shape global industries. Policymakers, meanwhile, must carefully balance data protection mandates with innovation incentives, ensuring that regulatory frameworks do not inadvertently stifle competition or innovation in less technologically advanced sectors. By focusing on the synergies between data security and overall business innovation, this paper provides a foundation for understanding how firms can thrive in an increasingly digital economy while navigating the risks and opportunities of data-intensive technologies.

References

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

Akey, Pat, Stefan Lewellen, and Inessa Liskovich, "Hacking Corporate Reputations," *SSRN Electronic Journal*, 01 2018.

Aldasoro, Iñaki, Leonardo Gambacorta, Paolo Giudici, and Thomas Leach, "The drivers of cyber risk," *Journal of Financial Stability*, Jun 2022, *60*, 100989.

Alderucci, Dean, Lee Branstetter, Eduard Hovy, Andrew Runge, and Nikolas Zolas, "Quantifying the impact of AI on productivity and labor demand: Evidence from U.S. Census microdata," *Working paper*, 2020.

Babina, Tania, Anastassia Fedyk, Alex He, and James Hodson, "Artificial intelligence, firm growth, and product innovation," *Journal of Financial Economics*, 2024, *151*, 103745.

_ , Saleem A Bahaj, Greg Buchak, Filippo De Marco, Angus K Foulis, Will Gornall, Francesco Mazzola, and Tong Yu, "Customer data access and fintech entry: Early evidence from open banking," Technical Report, National Bureau of Economic Research 2024.

Baker, Andrew C., David F. Larcker, and Charles C.Y. Wang, "How much should we trust staggered difference-in-differences estimates?," *Journal of Financial Economics*, 2022, *144* (2), 370–395.

Bena, Jan, Isil Erel, Daisy Wang, and Michael S Weisbach, "Specialized investments and firms' boundaries: Evidence from textual analysis of patents," Technical Report, National Bureau of Economic Research 2021.

Boasiako, Kwabena A. and Michael O'Connor Keefe, "Data breaches and corporate liquidity management," *European Financial Management*, 2021, *27* (3), 528–551.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess, "Revisiting Event Study Designs: Robust and Efficient Estimation," Forthcoming in ReStud, 2108.12419, arXiv.org 2022.

Callaway, Brantly and Pedro H. C. Sant'Anna, "Difference-in-Differences with Multiple Time Periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

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.

Crouzet, Nicolas and Janice Eberly, "Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles," Technical Report w25869, NBER 2019.

_ and _ , "Intangibles, Markups, and the Measurement of Productivity Growth," *Journal of Monetary Economics*, 2021.

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.

de Chaisemartin, Clément and Xavier d'Haultfœuille, "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 2020, *110* (9), 2964–2996.

Duffie, Darrell and Jeremy Younger, "Cyber Runs," Hutchins Center Unpublished Working Paper 51, Brookings Institution, Washington, D.C. 2019.

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, Dhruv Singal, Laura Veldkamp, and Venky Venkateswaran, "Valuing Financial Data," Working Paper w29894, National Bureau of Economic Research 2022.

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

Fedyk, Anastassia, Tatiana Fedyk, James Hodson, and Natalya V Khimich, "Do consulting services affect audit quality? Evidence from the workforce," *Working paper*, 2023.

Ganglmair, Bernhard, W Keith Robinson, and Michael Seeligson, "The rise of process claims: Evidence from a century of US patents," *Available at SSRN 4069994*, 2022.

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.

Goldstein, Itay, Chester S. Spatt, and Mao Ye, "The Next Chapter of Big Data in Finance," *The Review of Financial Studies*, 2024, *37* (1), 1–25.

Goodman-Bacon, Andrew, "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, 2021, *225* (2), 254–277.

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.

Hoberg, Gerard and Gordon M. Phillips, "Scope, Scale and Concentration: The 21st Century Firm," *Journal of Finance*, 2024. Forthcoming.

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

_ , _ , and Nagpurnanand Prabhala, "The Impact of Industry Fluidity on Firm Performance," *Journal of Financial Economics*, 2014, *117* (3), 483–512.

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

Huang, Henry and Chong Wang, "Do Banks Price Firms' Data Breaches?," *The Accounting Review*, 2021, *96* (3), 261–286.

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

Jiang, Wei, Yuehua Tang, Rachel (Jiqiu) Xiao, and Vincent Yao, "Surviving the Fintech Disruption," NBER Working Paper w28668, NBER 2021.

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

Kakhbod, Ali, Leonid Kogan, Peiyao Li, and Dimitris Papanikolaou, "Measuring Creative Destruction," *Available at SSRN 5008685*, 2024.

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.

Liu, Jinyu and Xiaoran Ni, "Ordeal by innocence in the big-data era: Intended data breach disclosure, unintended real activities manipulation," *European Financial Management*, 2023.

Mihet, Roxana and Thomas Philippon, "The economics of big data and artificial intelligence," *International Finance Review*, 2019, pp. 29–43.

Perkins Coie LLP, "Security Breach Notification Chart," https://www.perkinscoie.com/en/news-insights/security-breach-notification-chart.html 2023. Accessed on: 2023-06-29.

Rock, Daniel, "Engineering Value: The Returns to Technological Talent and Investments in Artificial Intelligence," *Management Science*, 2021, *67* (9), 5556–5580.

Scherbina, Anna and Bernd Schlusche, "The Effect of Malicious Cyber Activity on the U.S. Corporate Sector," Working Paper w29963, National Bureau of Economic Research 2022.

Sun, Liyang and Sarah Abraham, "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

Appendix A Variable descriptions

Table A.1: Variable Description

Variable	Description
Data risk score	Risk score built on the method developed by Florackis (2023) for US-based publicly listed firms. This variable quantifies data risk from 0 to 1. It is calculated through textual analysis of a firms annual 10-K filings, comparing the language on cyber risk-factors to that of the previous years filings from firms that suffered cyberattacks. A higher score suggests greater similarity and, thus, a higher risk of cybersecurity vulnerabilities.
Citation-weighted patent count	Sum of [one plus (cites / mean cites)] over all the patents filed by the firm in a given year. Where mean cites is the average number of cites for all the patents filed in the year. Thus, a patent with zero citations counts as one.
Cybersecurity patent	A patent classified by USPTO under any of the following CPC codes that relate to cybersecurity: G06F21/, H04L9/, H04l63/, G06F11/14, G06F12/14, H04L63/, 04W12/,G06Q20/382,H04B10/85, H04L2012/5687,H04M3/42008, G06Q50/265, H04L2209/42
Product patent	A patent which has majority of its claims classified as product claims by Ganglmair et al. (2022)
Process patent	A patent which has majority of its claims classified as process claims by Ganglmair et al. (2022)
AI-Intensive firm (Yes = 1)	Firms that file at least ten AI patents in the period 2000-2020, or those that are sufficiently close to the firms filing ten or more AI patents. AI patents are identified using data from Giczy et al. (2022). Firm closeness is identified using data from Hoberg and Phillips (2016).
In-house Cyber security firm (Yes = 1)	Firms that file both cybersecurity and non-cybersecurity patents and cite their own cybersecurity patents in other patents. This captures the set of firms building and using data security innovation themselves.
Assets	Compustat item at
R&D Expenditure	Compustat item xrd
Knowledge Assets	Item knowCapital in the dataset provided by Ewens et al. (2020). A measure of knowledge capital calcualted using R&D net of depreciation.
Market Equity	Market value of equity. Derived from Compustat items. Market Equity = $\text{prcc}_f * csho$
Tobin's Q	Derived from Compustat items. Tobin's Q = [at - ceq + Market Equity) / at
Leverage	Defined as long-term debt + debt in current liability as a share of total assets. Derived from compustat items. Leverage = (dltt +dlc) / at
Return on Assets (ROA)	Operating income before depreciation to total assets. Derived from Compustat items. ROA = oibdp / at
Asset Tangibility	Total property, plant and equipment to total assets. Derived from Compustat items. Tangibility = ppent / at
Book to market ratio	Book value of common equity (ceq) to Market Equity.
Cash to asset ratio (COA)	Cash holdings to assets. Derived from Compustat items. COA = che / at

Appendix B Empirical Strategy

Instrumental Variable: Data Breach Notification Laws

To address the potential endogeneity issues in our analysis of the impact of data risk on firm innovation and growth, we employ Data Breach Notification Laws (DBNLs) as an instrumental variable. DBNLs mandate organizations to notify affected individuals, regulatory authorities, and other stakeholders when a security breach involving unauthorized access, disclosure, or loss of personal data occurs. These laws serve to protect consumer privacy and enhance corporate accountability. The staggered adoption of DBNLs across different U.S. states provides a natural experiment setting, allowing us to exploit the exogenous variation

in data risk introduced by these laws.

Appropriateness of DBNLs as Instruments

DBNLs are appropriate instruments for several reasons:

- **Exogeneity**: The staggered implementation of DBNLs across states is driven by legislative processes and is unlikely to be correlated with individual firms' prior levels of innovation or specific data risk profiles. This staggered rollout introduces exogenous variation in data risk that is not directly influenced by firms' endogenous decisions, providing a robust source of exogenous variation.
- **Impact on Data Risk**: DBNLs significantly increase the cost and consequences of data breaches for firms, thereby directly affecting firms' data risk. This is evident from the increased efforts by firms to enhance data protection measures following the adoption of these laws, as documented in previous studies (Boasiako and Keefe, 2021; Liu and Ni, 2023).
- **Legal Mandate**: The legal requirement to disclose data breaches under DBNLs ensures that firms cannot underreport or hide incidents, leading to a more accurate and consistent measure of data risk across states and over time.

Potential Limitations

While DBNLs provide a powerful identification strategy, there are potential limitations to consider:

- **Multi-state Firms**: Firms operating in multiple states may be affected by DBNLs earlier than the state of their headquarters, potentially leading to spillover effects. This could contaminate the control group and attenuate the estimated impact of DBNLs. We mitigate this by focusing on firms headquartered in states with staggered DBNL adoption and by considering these spillover effects in our robustness checks.
- **Heterogeneity in Implementation**: Differences in the stringency and enforcement of DBNLs across states could lead to heterogeneous effects. While we control for state fixed effects and perform robustness checks, some variation in the impact of DBNLs may still exist.
- **Temporal Dynamics**: The nature of data risk and firms' responses to DBNLs may evolve over time. Our analysis accounts for these dynamics by including time fixed effects and examining the effects over different time horizons.

By addressing these potential limitations and explaining the appropriateness of DBNLs as instruments, we strengthen the validity of our empirical strategy and provide a more nuanced understanding of the causal impact of data risk on firm outcomes.

Appendix C Theoretical derivations

C.1 Model Solution Details

There are two sources of uncertainty in firm i's problem at date t: the (random) optimal technique $\theta_{i,t}$, and the aggregate price P_t. Let $(\hat{\mu}_{i,t}, \Omega_{i,t})$ denote the conditional mean and precision of firm i belief about $\theta_{i,t}$ given its information set at date t, $\mathcal{I}_{i,t}$.

In this section, we will first describe the firm belief updating process about its optimal technique. Next, we argue that in this environment, the firm's optimal production choice is deterministic, and thus the price is deterministic as well. Finally, we lay out the full set of equations that characterize the equilibrium of this economy with two groups of firms.

Belief updating The information problem of firm i about its optimal technique $\theta_{i,t}$ can be expressed as a Kalman filtering system, with a 2-by-1 observation equation, $(\hat{\mu}_{i,t}, \Omega_{i,t})$.

We start by describing the Kalman system, and show that the sequence of conditional variances is deterministic. Note that all the variables are firm specific, but since the information problem is solved firm-by-firm, for brevity we suppress the dependence on firm index i.

At time t, each firm observes two types of signals. First, date $t-1$ output provides a noisy signal about θ_{t-1}:

$$y_{t-1} = \theta_{t-1} + \epsilon_{a,t-1}, \tag{25}$$

where $\epsilon_{a,t} \sim \mathcal{N}(0,)$. We provide model detail on this step below. Second, the firm observes $n_t = z_t$ data points as a bi-product of its economic activity. The set of signals $\{s_{t,m}\}_{m\in[1:n_{i,t}]}$ are equivalent to an aggregate (average) signal $\bar{s}_t$ such that:

$$\bar{s}_t = \theta_t + \epsilon_{s,t}, \tag{26}$$

where $\epsilon_{s,t} \sim \mathcal{N}(0,)$. The state equation is

$$\theta_t - \bar{\theta} = \rho(\theta_{t-1} - \bar{\theta}) + \eta_t,$$

where $\eta_t \sim \mathcal{N}(0,)$.

At time, t, the firm takes as given:

$$\begin{aligned}\widehat{\mu}_{t-1} &= \mathbb{E}\big[\theta_t \mid s^{t-1}, y^{t-2}\big] \\ \Omega_{t-1}^{-1} &= Var\big[\theta_t \mid s^{t-1}, y^{t-2}\big]\end{aligned}$$

where $s^{t-1} = \{s_{t-1}, s_{t-2}, \dots\}$ and $y^{t-2} = \{y_{t-2}, y_{it-3}, \dots\}$ denote the histories of the observed variables, and $s_t = \{s_{t,m}\}_{m\in[1:n_{i,t}]}$.

We update the state variable sequentially, using the two signals. First, combine the

priors with y_{t-1}:

$$\begin{aligned}
\mathbb{E}\left[\theta_{t-1} \mid \mathcal{I}_{t-1}, y_{t-1}\right] &= \frac{\Omega_{t-1}\widehat{\mu}_{t-1} + y_{t-1}}{\Omega_{t-1}+} \\
V\left[\theta_{t-1} \mid \mathcal{I}_{t-1}, y_{t-1}\right] &= \left[\Omega_{t-1} + \right]^{-1} \\
\mathbb{E}\left[\theta_{t} \mid \mathcal{I}_{t-1}, y_{t-1}\right] &= \bar{\theta} + \rho \cdot \left(\mathbb{E}\left[\theta_{t-1} \mid \mathcal{I}_{t-1}, y_{t-1}\right] - \bar{\theta}\right) \\
V\left[\theta_{t} \mid \mathcal{I}_{t-1}, y_{t-1}\right] &= \rho^2\left[\Omega_{t-1} + \right]^{-1} +
\end{aligned}$$

Then, use these as priors and update them with $\bar{s}_t$:

$$\widehat{\mu}_t = \mathbb{E}\left[\theta_t \mid \mathcal{I}_t\right] = \frac{\left[\rho^2\left[\Omega_{t-1} + \right]^{-1} + \right]^{-1} \cdot \mathbb{E}\left[\theta_t \mid \mathcal{I}_{t-1}, y_{t-1}\right] + \bar{s}_t}{\left[\rho^2\left[\Omega_{t-1} + \right]^{-1} + \right]^{-1} +} \tag{27}$$

$$\Omega_t^{-1} = Var\left[\theta \mid \mathcal{I}_t\right] = \left\{\left[\rho^2\left[\Omega_{t-1} + \right]^{-1} + \right]^{-1} + \right\}^{-1} \tag{28}$$

Multiply and divide equation (27) by Ω_t^{-1} as defined in equation (28) to get

$$\widehat{\mu}_t = (1 - n_t\sigma_\epsilon^{-2}\Omega_t^{-1})\left[\bar{\theta}(1-\rho) + \rho\left((1-M_t)\mu_{t-1} + M_t\tilde{y}_{t-1}\right)\right] + n_t\sigma_\epsilon^{-2}\Omega_t^{-1}\bar{s}_t, \tag{29}$$

where $M_t = \sigma_a^{-2}(\Sigma_{t-1} + \sigma_a^{-2})^{-1}$.

Equations (28) and (29) constitute the Kalman filter describing the firm dynamic information problem. Importantly, note that Ω_t^{-1} is deterministic.

C.2 Modeling quadratic-normal signals from output

When y_{t-1} is observed, agents can back out A_{t-1} exactly. To keep the model simple, we assumed that when agents see A_{t-1}, they also learn whether the quadratic term $(a_{t-1} - \theta_{t-1} - \epsilon_{a,t-1})^2$ had a positive or negative root. An interpretation is that they can figure out if their action a_t was too high or too low.

Relaxing this assumption complicates the model because, when agents do not know which root of the square was realized, the signal is no longer normal. One might solve a model with binomial distribution over two normal variables, perhaps with other simplifying assumptions. For numerical work, a good approximate solution would be to simulate the binomial-normal and then allows firms to observe a normal signal with the same mean and same variance as the true binomial-normal signal. This would capture the right amount of information flow, and keep the tractability of updating with normal variables.

C.3 The cybersecurity planning problems: optimality conditions and steady state results

C.3.1 H-type firm

The current-value Hamiltonian function for the H-type firm:

$$H(\Omega_{H,t};\tau_t;\delta_t;p_{H,t}) = \Pi_{H,t,\text{sec}}$$
$$+\beta p_{H,t+1}\left\{\left[\rho^2(\Omega_{H,t}+\sigma^{-2})^{-1}+\sigma^2\right]^{-1}+\left[1-\vartheta e^{-\left(\tau_t-\iota\frac{1-u}{u}\delta_t\right)}\right]z\sigma^{-2}-\Omega_{i,t}\right\} \tag{30}$$

where $p_{H,t}$ is the shadow-price or co-state variable associated with the state variable. The transversality condition is $\lim_{t\to\infty}\Omega_{H,t}\beta^t p_{H,t}=0$.

The first-order optimality conditions:

$$\frac{\partial H}{\partial \tau_t}=0\Rightarrow \beta p_{H,t+1}=\frac{1-b\overline{A}e^{b\left(\tau_t-\iota\frac{1-u}{u}\delta_t\right)}}{\vartheta e^{-\left(\tau_t-\iota\frac{1-u}{u}\delta_t\right)}z\sigma^{-2}} \tag{31}$$

$$\frac{\partial H}{\partial \delta_t}=0\Rightarrow \beta p_{H,t+1}=\frac{\pi_t-b\overline{A}\iota e^{b\left(\tau_t-\iota\frac{1-u}{u}\delta_t\right)}}{\vartheta\iota e^{-\left(\tau_t-\iota\frac{1-u}{u}\delta_t\right)}z\sigma^{-2}} \tag{32}$$

$$\beta p_{H,t+1}-p_{H,t}=-\frac{\partial H}{\partial \Omega_{H,t}}\Rightarrow\left[\rho+\frac{\sigma^2}{\rho}(\Omega_{H,t}+\sigma^{-2})\right]^{-2}\beta p_{H,t+1}=p_{H,t}-\Omega_{H,t}^{-2} \tag{33}$$

From (31) and (32), it emerges a constant optimal trading price, which is simply $\pi_t=\iota$. The price of protection is directly associated with the degree of its own nonrivalry. If protection is completely non-rival (i.e., $\iota=0$), then its price is zero; if protection is fully rival, its price is 1.

Replacing (31) into (33), and evaluating in the steady state, one gets:

$$\Gamma_H=\frac{\vartheta z e^{-\left(\tau^*-\iota\frac{1-u}{u}\delta^*\right)}}{1-b\overline{A}e^{b\left(\tau^*-\iota\frac{1-u}{u}\delta^*\right)}}, \tag{34}$$

with Γ_H defined as $\Gamma_H\equiv\left\{\frac{1}{\beta}-\left[\rho+\frac{\sigma^2}{\rho}(\Omega_H^*+\sigma^{-2})\right]^{-2}\right\}(\Omega_H^*)^2\sigma^2$.

Given constraint (15), it is also true, for the H-firms:

$$\Xi_H=\left[1-\vartheta e^{-\left(\tau^*-\iota\frac{1-u}{u}\delta^*\right)}\right]z, \tag{35}$$

with $\Xi_H\equiv\left\{\Omega_H^*-\left[\rho^2(\Omega_H^*+\sigma^{-2})^{-1}+\sigma^2\right]^{-1}\right\}\sigma^2$.

Combining expressions (34) and (35), one obtains a steady state relation that allows for the derivation of Ω_H^*:

$$\Gamma_H=\frac{z-\Xi_H}{1-b\overline{A}\left(\frac{\vartheta z}{z-\Xi_H}\right)^b} \tag{36}$$

Γ_H is such that if $\Omega_H^*=0$ then $\Gamma_H=0$ and if $\Omega_H^*\to+\infty$ then $\Gamma_H\to+\infty$.

Ξ_H is such that if $\Omega_H^* = 0$ then $\Xi_H = -\frac{1}{1+\rho^2}$ and if $\Omega_H^* \to +\infty$ then $\Xi_H \to +\infty$. Hence, if $\Omega_H^* = 0$ then $\frac{z-\Xi_H}{1-b\overline{A}\left(\frac{\vartheta z}{z-\Xi_H}\right)^b} = \frac{z+\frac{1}{1+\rho^2}}{1-b\overline{A}\left(\frac{\vartheta z}{z+\frac{1}{1+\rho^2}}\right)^b}$; this is a positive value for $b\overline{A}\left(\frac{\vartheta z}{z+\frac{1}{1+\rho^2}}\right)^b < 1$. If $\Omega_H^* \to +\infty$ then $\frac{z-\Xi_H}{1-b\overline{A}\left(\frac{\vartheta z}{z-\Xi_H}\right)^b} \to -\infty$.

By combining the above reasoning, as long as $b\overline{A}\left(\frac{\vartheta z}{z+\frac{1}{1+\rho^2}}\right)^b < 1$, the l.h.s. of (36) (positively sloped) will intersect the r.h.s. of (36) (negatively sloped) at one single point, and therefore a unique Ω_H^* is derived.

Thus, condition $b\overline{A}\left(\frac{\vartheta z}{z+\frac{1}{1+\rho^2}}\right)^b < 1$ must hold, which can be rewritten as a constraint on ϑ: $\vartheta < \frac{z+\frac{1}{1+\rho^2}}{z}\left(b\overline{A}\right)^{-1/b}$. Because $\vartheta \leq 1$, this constraint is always satisfied as long as $b\overline{A} < 1$.

From (35) also note that the value of security that firm H holds after trade is also a unique constant value,

$$\tau^* - \iota\frac{1-u}{u}\delta^* = -\ln\left(\frac{z-\Xi_H}{\vartheta z}\right) \tag{37}$$

C.3.2 L-type firm

Turning to the L-type firm, the current-value Hamiltonian is:

$$H(\Omega_{L,t};\delta_t;p_{L,t}) = \Pi_{L,t,\text{sec}}$$
$$+\beta p_{L,t+1}\left\{\left[\rho^2(\Omega_{L,t}+\sigma^{-2})^{-1}+\sigma^2\right]^{-1} + \left(1-\vartheta e^{-\delta_t}\right)z\sigma^{-2} - \Omega_{i,t}\right\} \tag{38}$$

The transversality condition: $\lim_{t\to\infty}\Omega_{L,t}\beta^t p_{L,t} = 0$.

The first-order conditions are:

$$\frac{\partial H}{\partial \delta_t} = 0 \Rightarrow \beta p_{L,t+1} = \frac{\pi_t}{\vartheta e^{-\delta_t} z\sigma^{-2}} \tag{39}$$

$$\beta p_{L,t+1} - p_{L,t} = -\frac{\partial H}{\partial \Omega_{L,t}} \Rightarrow \left[\rho + \frac{\sigma^2}{\rho}(\Omega_{L,t}+\sigma^{-2})\right]^{-2}\beta p_{L,t+1} = p_{L,t} - \Omega_{L,t}^{-2} \tag{40}$$

Replace (39) into (40), and recall that we already know that $\pi_t = \iota$. With this information, the following steady state condition holds:

$$\Gamma_L = \vartheta z e^{-\delta^*}, \tag{41}$$

with $\Gamma_L \equiv \left\{\frac{1}{\beta} - \left[\rho + \frac{\sigma^2}{\rho}(\Omega_L^* + \sigma^{-2})\right]^{-2}\right\}(\Omega_L^*)^2\sigma^2$.

Given constraint (17),

$$\Xi_L = \left(1-\vartheta e^{-\delta^*}\right)z, \tag{42}$$

with $\Xi_L \equiv \left\{\Omega_L^* - \left[\rho^2(\Omega_L^*+\sigma^{-2})^{-1}+\sigma^2\right]^{-1}\right\}\sigma^2$.

From (41) and (42), a simple expression emerges for the determination of Ω_L^*,

$$\Gamma_L = z - \Xi_L \tag{43}$$

Equation (43) allows for the derivation of a unique Ω_L^*, because the l.h.s. of the expression is a continuous increasing function starting at zero and diverging to infinity (as Ω_L increases) and the l.h.s. is a continuous decreasing function starting at a positive value and falling to minus infinity (as Ω_L increases).

From (42), one can also compute the steady state value of the amount of security bought by firm L:

$$\delta^* = -\ln\left(\frac{z - \Xi_L}{\vartheta z}\right) \tag{44}$$

A unique δ^* exists as well.

By now, we have computed all the relevant steady state values: Ω_H^* and Ω_L^*, and also δ^* (determined from the L-firm problem), and τ^*, determined from (37) after knowing δ^* (the H-type only decides how much to invest in cyberprotection after knowing how much protection firms in the L sector are willing to buy at price $\pi_t = \iota$).

C.3.3 Steady-state

Possible steady state scenarios:

- (*i*) The cybersecurity optimal result is such that $\tau^* \leq 0$: firms H do not invest in cybersecurity $\tau^* = 0$ and firms L have no cyberprotection to buy, $\delta^* = 0$. Firms face the problem with no security and their profits are: $\Pi_{H,cy}^* = \Pi_{L,cy}^*$.

- (*ii*) The cybersecurity optimal result is such that $\tau^* > 0$, $\delta^* \leq 0$: firms L will not buy any protection and face the no-protection problem, with profits $\Pi_{L,cy}^*$. Firms of the H type have two possibilities: to invest τ^*, even though they cannot optimally exchange protection, or not to invest; they compare profits $\Pi_{H,\text{sec}}^*$ and $\Pi_{H,cy}^*$ and choose the option that delivers the highest profits.

- (*iii*) The cybersecurity optimal result is such that $\tau^* > 0$ and $\delta^* > 0$: firms find it optimal to invest a positive value in cybersecurity (H) and to trade a positive amount of cybersecurity. In this case, the best option is the cybesecurity one with profits $\Pi_{H,\text{sec}}^*$ and $\Pi_{L,\text{sec}}^*$

Note that conditions $\tau^* > 0$ and $\delta^* > 0$ impose relevant constraints on parameter values, namely, in the first case, $z > \Xi_H$ and $\vartheta > \frac{z-\Xi_H}{z}$ and, in the second case, $z > \Xi_L$ and $\vartheta > \frac{z-\Xi_L}{z}$. These results suggest that investment and trading in cybersecurity require the cybercrime index ϑ to be above a given threshold.

C.4 Comparative statics

A few intuitive comparative statics outcomes (in the cybersecurity setting, i.e., for $\tau^* > 0$, $\delta^* > 0$):

- (i) $\Delta z > 0$: l.h.s. of (36) does not shift; r.h.s. of (36) shifts right $\Rightarrow$ higher Ω_H^* / l.h.s. of (43) does not move; r.h.s. of (43) shifts right $\Rightarrow$ higher Ω_L^* / δ^* and τ^* increase / output of both types of firms will increase.

- (ii) $\Delta u > 0$: Ω_H^*, Ω_L^*, and δ^* do not change; only τ^* decreases - logical result: relatively more firms investing in cyberprotection implies lower investment by each of them to attain the optimal result. Output of L firms is maintained; output of H firms is also maintained (the decrease in τ^* is compensated by the increase in u and, according to (37), there is no change on the available protection and, thus, on output).

- (iii) $\Delta \iota > 0$: Ω_H^*, Ω_L^*, and δ^* do not change; only τ^* decreases - logical result: a lower degree of non-rivalry in selling protection implies H firms will invest more to keep more protection and to profit more from trading. Output does not change for any of the firms for reasons similar to those of the previous item.

- (iv) $\Delta \vartheta > 0$: l.h.s. of (36) does not shift; r.h.s. of (36) shifts right $\Rightarrow$ higher Ω_H^* (this is the positive effect that innovation from cybersecurity has over knowledge when H firms increase cybersecurity in response to cybercrime) / Ω_L^* remains unchanged / δ^* increases (L firms demand more security to face higher risks) / τ^* increases due to the increase on δ^* and directly on ϑ. Output levels will increase, given the corresponding expressions.

- (v) $\Delta b > 0$: l.h.s. of (36) does not shift; r.h.s. of (36) shifts right $\Rightarrow$ higher Ω_H^* / Ω_L^* remains unchanged / τ^* increases because of the increase in Ω_H^*; δ^* does not change / the output of L firms does not change / the output of H firms increases.

C.5 Simulating the economy

Take the values in the table below.

Parameter	Symbol	Value
Data endowment	z	10
Coefficent of the AR(1) process	ρ	0.9
Variances	σ^2	2.5
Share of H-type firms	u	1/3
Non-rivalry parameter	ι	0.6
Intertemporal discount factor	β	0.96
Data risk index	ϑ	0.75
Maximum quality	$\overline{A}$	25
Inonvation externality	b	0.035
Capital cost	r	1

For these parameters: $\Omega_H^* = 3.224$ and $\Omega_L^* = 1.609$. These results are found in the intersection of the l.h.s. and r.h.s. of (36) and (43) in Theory **[Fig.1]**.

Applying the corresponding formulas, $\delta^* = 0.130$ and $\tau^* = 1.296$ (these are both positive values and, therefore, firms engage in data security investment and data security trading).

Replacing the equilibrium values in the expressions for output and profits, $A_H^* = 23.207$ and $A_L^* = 21.879$ ($A_H^* > A_L^*$); $\Pi_{H,\text{sec}}^* = 21.068$ and $\Pi_{L,\text{sec}}^* = 20.800$ ($\Pi_{H,\text{sec}}^* > \Pi_{L,\text{sec}}^*$). Also, $Y^* = uA_H^* + (1-u)A_L^* = 22.321$.

C.6 Comparative statics

How do steady state values change with data risk?

Recall that $\vartheta \in [0,1]$. Evaluating the model for different values of ϑ (and letting all other values be as in Table 1), we find two thresholds: at $\vartheta = \frac{z-\Xi_L}{z} = 0.6583$, optimal security purchasing, δ^*, changes from negative to positive, implying that firms L buy protection only for $\vartheta > 0.6583$. For $\vartheta \leq 0.6583$, H firms have to choose whether to invest in protection or not, knowing that they will sell no protection. They compare profits $\Pi_{H,\text{sec}}^*$ and $\Pi_{H,cy}^*$; these are equal around $\vartheta = 0.3$. For $\vartheta > 0.3$, H-type firms invest in protection, otherwise they do not.

Theory Fig. 15 draws profits without protection for both firms (these are identical), the profits of the H firms with security investment, and the profits of the L firms under security trading. The two mentioned thresholds are highlighted.

Hence: for $\vartheta \leq 0.3$, H-firms do not invest in data protection and L-firms do not buy protection; for $0.3 < \vartheta \leq 0.6583$, H firms invest in protection and L firms buy no protection; for $\vartheta > 0.6583$, H-type firms invest in protection and L-type firms buy protection. In this last segment, the higher the value of ϑ, the more the H firms invest and the more L firms buy.

Theory Fig. 16 presents the investment and trading levels. Again, the two thresholds are clear (notice the second jump in τ^*; this occurs because to the right of that point, H-type firms need to invest in security for their one use but also to sell to firms in the L group).

Theory Fig. 17: output of each type of firm and aggregate output, for different levels of data risk. In the first segment, the output is the same (the firms are identical); in the second segment, L firms face increasing risk but do not protect and, consequently, output falls (because the stock of knowledge falls); H firms start investing in data security what has the innovation side effect and, therefore, they are able to increase output. In the third segment, H firms continue to invest in data protection and innovate; L firms start purchasing security that they cannot use to innovate but that prevents output from falling (i.e., it allows to maintain the stock of knowledge as the data risk increases).

The aggregate output is a weighted average of the output of the two types of firms (recall that, in the example, L firms are two thirds of the total number of firms). Concerning the evolution of Y^* as ϑ increases, one notices that an initial fall is counteracted when H firms start to invest in protection, and this process gains a new impetus when L firms start protecting as well.

We can recover the representation of profits in Theory Fig.2, to draw the actual profits, given the choices of firms on whether to get protection or not.

Theory Fig. 18 clarifies again the existence of three stages and the fact that cyber crime is much less harmful for H-type firms, because these make use of the innovation externality that data security allows for.

C.7 Does data growth cause economic growth?

In the model, there are various parameters whose values can change - $\vartheta, \iota, u, b, ...$ - but only one can grow in a sustained way over time, which is the endowment of data, z. The question is: if one makes z to increase over time at a constant rate, will the economy's output also grow over time at a constant rate?

The answer is no: simulations show that although the increase in z leads to increases in Ω_H^* and Ω_L^*, they also lead to falls in δ^* and τ^* (more data and a same data risk lead to the need of less protection). For large values of z, τ^* becomes zero, and without investment in cybersecurity there is no data risk induced innovation and the maximum quality of output cannot expand. The increases in Ω_H^* and Ω_L^* are associated with decreasing marginal returns and, therefore, although z might grow in a sustained way, this is not accompanied by an increase in the firms' output.