Data as a Networked Asset*

Bo Bian† Qiushi Huang‡ Ye Li§ Huan Tang¶

April 18, 2025

Data is non-rival: a firm's data can be used simultaneously by others, and information about its customers benefits other firms even across industries. How is data being shared? Using granular information on mobile app usage, functionalities, and connections with data analytics platforms, we uncover a network of inter-firm data flows. Data sharing generates comovements in operational, financial, and stock-market performances among data-connected firms, beyond what traditional economic linkages can explain, and induces strategic complementarity in firms' product-design choices. Apple's App Tracking Transparency policy, which restricts inter-firm data flows, weakens these patterns, providing causal evidence of the role of data sharing. To explain these findings, we develop a dynamic network model of data economy, where firm growth becomes interconnected through data sharing. The model introduces a network-augmented Gordon growth formula to value data-generated cash flows, capturing direct and indirect network externalities over multiple time horizons. Our metrics of valuation centrality identify systemically important firms that disproportionately influence the data economy due to their pivotal positions within the data-sharing network.

Keywords: Data, customer capital, intangible, network, systemic importance, cyberattack

JEL Codes: D62, D85, E22, E23, G12, G14, L51, L86, O33

*We thank Cecilia Bustamante, Philip Bond, Tony Cookson, Nicolas Crouzet, Jerry Hoberg, Allen Hu, Adriano Rampini, Amit Seru, Luke Taylor, and participants at the Maryland Junior Finance Conference. RCFS Winter Finance Conference and UNC-Duke Corporate Finance Conference for their helpful comments and suggestions. We thank Yiming Ma, Alina Song, and Bolin Xu for outstanding research assistance. This research received financial support from the Stevens Center for Innovation in Finance at the Wharton School of the University of Pennsylvania. First version: November 18, 2024.

†University of British Columbia, Sauder School of Business, bo.bian@sauder.ubc.ca.

‡Shanghai Advanced Institute of Finance, qshuang@saif.sjtu.edu.cn.

§University of Washington, Foster School of Business, liye@uw.edu.

¶University of Pennsylvania, the Wharton School, huan.ht.tang@gmail.com

1 Introduction

Data has emerged as a highly productive asset. It is non-rival: one firm's use of data does not diminish its availability for others, allowing data to be utilized by multiple firms simultaneously (Jones and Tonetti, 2020). Furthermore, data exhibits externalities: information collected by one firm can benefit other firms. Data on a firm's customers can be useful for profiling other firms' customers by revealing underlying economic forces or as training data for prediction models (Choi et al., 2019; Ichihashi, 2021; Acemoglu et al., 2022). Data non-rivalry, combined with externality, significantly expand the potential uses of a firm's data, often transcending industry boundaries.

In this paper, we explore several questions: How is data collected by one firm shared with others? What is the scope of data sharing? What are the economic implications, particularly, how data sharing affects firms' decision-making and propagates shocks across firms? How do privacy regulations influence the economics of data sharing, and what unintended consequences might arise? Finally, we investigate which firms are systemically important in the data economy. These firms likely contribute significant amounts of data, but their systemic importance cannot be simply determined by size alone; the network topology of data flows across firms plays a critical role.

We trace inter-firm data flows originating from mobile applications (apps), which have become the primary channels for data collection in the economy. Our sample contains 1,031 app-owning public firms that account for more than 60% of the total assets of Computat firms. Firms may operate one or multiple mobile apps. These apps collect data and transmit it to Software Development Kits (SDKs) specializing in data aggregation and analytics. A crucial function of SDKs is merging information from various apps associated with the same consumer, creating comprehensive customer profiles that can be utilized by multiple firms and across industries. By sharing data with SDKs, firms gain access to these valuable signals for customer profiling in return.

Using granular information on app-level SDK installations, we construct measures of data

connectedness based on firms' overlap in SDK usage. Specifically, two firms are considered connected when a common set of data-analytics SDKs are installed on their apps, with the degree of connectedness increasing in the SDK overlap. This approach enables us to construct the first measure of inter-firm data sharing and to map out the network structure of data flows.

The signals that connected firms receive from SDKs contain data from one another. When one firm gains more customers, it collects more data, which is shared through the SDKs with its connected firms, making them more informed about their customers as well. Being able to profile customers more effectively in turn allows these connected firms to acquire more customers and improve revenue generation. Our empirical analysis confirms this dynamic. We find that data sharing generates comovements in firms' operational performances. The economic magnitude is large, more than doubling that of product similarity measure from Hoberg and Phillips (2016) in explaining the comovement in operational performances. Additionally, data-connected firms exhibit stock-return correlations that cannot be explained by standard asset pricing risk factors or other common exposures, such as product overlap or common analyst coverage.

Our paper is the first to uncover this new form of economic linkage. Recently, with rising consumer awareness of privacy concerns, data privacy regulations, such as GDPR in Europe and CCPA in the U.S. (California), have imposed strict rules on data collection and sharing. We find that one unintended consequence of these policies is a weakening of performance comovement between data-connected firms. In a difference-in-differences framework, we explore the introduction of Apple's user privacy framework, App Tracking Transparency (ATT), as the policy shock. By restricting firms' ability to share data and limiting SDKs' capacity to merge data from different firms into unified customer profiles, ATT curtails inter-firm data flows. We find that performance comovement induced by data connectedness significantly weakened after ATT.

By examining the impact of ATT, we not only shed light on the consequences of upcoming

privacy regulations from both private and public sectors but also validate that cross-firm overlap in data-analytics SDKs is fundamentally about data sharing. Importantly, ATT does not influence other forms of firm overlap, such as product overlap, that contribute to performance comovements.

Interconnectedness implies that shocks to one firm propagate to others, which in turn contributes the comovement in firms' performances. Using cyberattacks as our empirical setting, we provide direct evidence of how data sharing facilitates shock propagation. Intuitively, a cyberattack reduces the focal firm's data stock, impairs its ability to collect data, and more broadly diminishes its operational efficiency that underpins customer engagement and data acquisition. We find that the local firm's data-connected peers experience significantly larger deterioration in operational and stock-market performance than firms not connected through data sharing or connected through other economic linkages. These findings underscore the importance of examining the network structure of data sharing to trace the ripple effects of cyberattacks and other shocks.

Motivated by the evidence on data sharing, we develop a dynamic network model of data economy that replicates the data-induced comovements in firms' performances and the ATT impact. In our model, firms accumulate raw data from customer activities and then share data through a network. Firm i is connected to firm j when its composite signal on customers depends on firm j's data, represented by the loading γ_{ij}. An increase in γ_{ij} leads to stronger performance comovements between firms i and j. These pairwise data connections are set to their empirical counterparts, i.e., the overlap of firms' data-analytics SDKs as measured from the data.

For each firm, the composite signal for customer profiling plays two roles. First, it helps the firm generate customer engagement, which in turn translates into an increase in customer capital and product demand.[1] Moreover, more customer engagement creates more data, leading to self-perpetuating data growth as in Farboodi and Veldkamp (2021). Importantly, the additional data

[1] One channel of customer capital build-up is that data improves firm-consumer match (Gourio and Rudanko, 2014).

also feeds into other firms' signals and contributes to their build-up of customer capital. Therefore, through data collection and sharing, the growth of data and customer capital becomes interconnected across firms, generating a channel of comovement and persistent shock propagation.

The second role of the composite signal relates to a critical intertemporal trade-off faced by firms. In our model, firms can adjust their product designs to prioritize either monetization or customer engagement. For instance, a software company might generate higher profits per interaction with customers by moving more features behind paywall, but this approach reduces the overall level of customer interaction. While this appears to be a static trade-off between intensive and extensive margins, it has dynamic implications: prioritizing monetization reduces customer engagement and thereby limits data accumulation. The composite signal mitigates this negative impact. Intuitively, when a firm has a deeper understanding of its customers, it can design products that emphasize profitability but still maintain customer engagement; likewise, if the firm prioritizes customer engagement and data collection—such as by offering more features for free—the negative impact on profitability is mitigated by the composite signal that enables the firm to charge higher prices for premium features, effectively capturing value from more engaged users. In summary, the composite signal for customer profiling alleviates the tension in the intertemporal trade-off between current profitability and data accumulation for the self-perpetuating growth.

This mechanism gives rise to intriguing product-design dynamics. When one firm prioritizes customer engagement, the data it generates enables its connected firms to boost customer engagement without significantly compromising profitability. As a result, the connected firms choose to stimulate customer engagement as well and collect more data. Conversely, when a firm prioritizes monetization, its reduced data collection and diminished data spillover makes it harder for the connected firms to balance profitability and customer engagement. For any given level of profitability, these connected firms have to sacrifice more customer engagement and collect less data as well.

In equilibrium, firms' product-design decisions exhibit "herding" behavior. Empirically, we find that a firm's product-design choices are strongly influenced by those of its data-connected peers, even after accounting for other common exposures, such as product overlap. Furthermore, herding in product design among data-connected firms is significantly weakened by the introduction of ATT, indicating that this empirical pattern is driven by data sharing.[2]

In our model, data functions as productive capital, analogous to the role of capital in classic investment theories (Hayashi, 1982; Abel and Eberly, 1994), with a firm's product-design choices mirroring investment decisions. Specifically, the marginal q of a firm's data—the derivative of its value function with respect to its data stock—drives the decision on whether to prioritize customer engagement and data collection in product design. However, there are two critical distinctions. First, data investment features a positive externality, in contrast to the traditional investment dynamics where one firm's investment often crowds out others' investment.[3] Second, firms' investment decisions are directly interconnected through data sharing in our model. A firm's data marginal q incorporates the expected trajectories of data inflows from other firms (indegree network externality) but disregards the data outflows to other firms (outdegree network externality). This internalization of indegree externality creates strategic complementarity, or herding behavior, among firms. Meanwhile, the failure to internalize outdegree externality leads to under-investment.

We derive a network-augmented Gordon growth formula for valuing data-driven firms. In our model, a firm's valuation in equilibrium, i.e., the present value of cash flows, is a function of its own data stock and data stocks of its peers that are connected via data sharing. We show that in the absence of data sharing, firms' valuations reduce to the standard Gordon growth formula, but under data sharing, firm-level growth in valuation is replaced by a firm's loading on the "community

[2]We focus on how a nonfinancial firm's incentive to generate data depends on other firms' choices. Farboodi and Veldkamp (2020) study how a trader produces different types of data depends on other traders' information choices.

[3]For example, investment by one firm may increase the cost of investment inputs and cost of financing or intensifying product-market competition that other firms face (e.g., Asriyan et al., 2024).

growth" of the entire data economy. Each firm's valuation can be decomposed into contributions from its own data and those from the data of its connected peers.[4]

Our model and the implied valuation formula provide a framework for quantifying firms' systemic importance. So far, our empirical findings on the comovement in data-connected firms' performances and their herding in product design are based on direct connections. This reduced-form strategy omits two critical aspects of network externality, the higher-order externalities and persistent impact over time. Impact of a firm's data on directly connected firms is likely to transmit further to their connected peers, resulting in higher-order cascading effects. Additionally, since data generates self-reinforcing growth as in Farboodi and Veldkamp (2021), a firm's data has persistent effects on itself and its connected peers by affecting the entire trajectory of data growth.

Based on our calibrated model, we develop metrics of firms' systemic importance that incorporates data spillover effects through both direct and indirect connections and over multiple time horizons. A key input is the data-sharing network that we measure directly from our sample.

Specifically, we solve for the present value of aggregate cash flows in the economy (aggregate valuation) as a function of all firms' data stocks (the state variables). This aggregate valuation captures all pathways of network propagation, accounting for the impact of any firm's data on other firms in the economy and on aggregate cash flows across all time horizons.[5] We decompose aggregate valuation into individual firms' contributions and demonstrate that a firm's contribution to aggregate valuation can differ significantly from its own valuation. For instance, before the introduction of ATT, Meta (formerly Facebook) had a ratio of aggregate valuation contribution to its own valuation of 2.5, indicating that removing Meta from the data economy would result in a loss of aggregate cash flows equivalent to 2.5 times Meta's own valuation. After ATT, this ratio de-

[4]Our model sheds light on the scope of data usage. Crouzet et al. (2024) emphasize the scope of intangible capital usage within firms. Our findings on data sharing reveals cross-firm usage of data as a particular type of intangibles.

[5]Veldkamp (2023) provides an overview on the current methods of valuing data.

clined to 1.2. These findings highlight that under data sharing, certain firms become systemically important to the data economy in ways that their market value or size alone does not fully capture.

Literature. To the best of our knowledge, we are the first to systematically characterize the complex network of data flows among public firms. Our research demonstrates that firms are not only linked by conventional economic relationships—such as product market overlap (Hoberg and Phillips, 2010, 2016, 2018), supply chain connections (Cohen and Frazzini, 2008; Menzly and Ozbas, 2010), shared analyst coverage (Ali and Hirshleifer, 2020), geographic exposure (Parsons et al., 2020), or technological proximity (Bloom et al., 2013; Liu and Ma, 2021)—but also by the exchange and use of data. Our contribution lies in documenting how this new type of economic linkage profoundly impacts firm valuation and stock comovement in the digital era.

Our paper contributes to the literature on the non-rival nature of intangible capital (e.g., McGrattan and Prescott, 2009, 2010; Varian, 2018; Jones and Tonetti, 2020; Cong et al., 2021; Crouzet et al., 2022; Liu et al., 2023) and is most related to Crouzet et al. (2024). Crouzet et al. (2024) study the scope of intangible capital usage and highlight both the positive and negative effects of broadening the scope on economic growth. In our paper, the scope of intangible (data) usage is reflected in the data-sharing network. While Crouzet et al. (2024) emphasize the within-firm usage of intangibles, our focus is on cross-firm data sharing. We directly measure the scope of data usage and embed the network structure in an otherwise canonic model of data-driven firm growth (e.g., Begenau et al., 2018; Jones and Tonetti, 2020; Farboodi and Veldkamp, 2021). Broadening the scope of data usage benefits growth. Reducing the scope, for example, through regulations on cross-firm data sharing, dampens growth and the valuation of data-driven firms in line with the empirical findings on the ATT impact (Bian et al., 2021).[6] We also highlight that data sharing facilitates growth but also propagates shocks by synchronizing firms' behavior, thus amplifying aggregate volatility.

[6]Bian et al. (2021) document a -3% cumulative abnormal return of data-reliant firms in the month following ATT.

This harmful effect differs from that in Crouzet et al. (2024) who connect an increase in the scope of intangible usage with a reduction in entrepreneurship. Based on our findings, an unintended consequence of privacy regulations is to reduce the comovement of firms' performances, which in turn moderates the aggregate fluctuation of data economy.

Our paper also contributes to the literature on measuring and valuing intangible capital (e.g., Eisfeldt and Papanikolaou, 2013; Gourio and Rudanko, 2014; Kogan et al., 2017; Peters and Taylor, 2017; Kelly et al., 2021; Bhandari and McGrattan, 2021; Dou et al., 2021; Ewens et al., 2024). How to value data as a productive asset has become an increasingly important question (Veldkamp, 2023). We highlight the limitations of cost-based methods for valuing data assets and emphasize the importance of incorporating the cross-firm scope of data usage in the valuation framework. The value of data assets—defined as the present value of data-generated cash flows—depends significantly on the extent to which data is shared and utilized within the whole data economy. Previous work on measuring data also takes an indirect approach by focusing on firms' decisions to obtain complementary labor inputs (Abis and Veldkamp, 2023; Corhay et al., 2024) or the outcome of data usage (e.g., Farboodi et al., 2024; Eeckhout and Veldkamp, 2022).

In our model, firms' decision-making and valuation are interconnected through a network adjacency matrix of data sharing, and the equilibrium conditions map to a spatial structure (de Paula, 2017) as in models of social connections (e.g., Glaeser and Scheinkman, 2000; Ballester et al., 2006; Graham, 2008; Calvó-Armengol et al., 2009; Bramoullé et al., 2009; Blume et al., 2015; Fogli and Veldkamp, 2021) Spatial models have been recently adopted in financial economics (Cohen-Cole et al., 2014; Ozdagli and Weber, 2017; Herskovic, 2018; Herskovic et al., 2020; Jiang and Richmond, 2021; Eisfeldt et al., 2022, 2023; Li et al., 2023). Our paper is the first to analyze the spatial structure of data flows. To characterize data accumulation over time, our model is fully stochastic and dynamic. Following Diebold and Yılmaz (2014), Ballester et al. (2006), and

Denbee et al. (2021), we decompose aggregate valuation of cash flows into firms' contributions and develop the first metric of firm systemic importance in the data economy.[7]

On the empirical side, by examining the impact of data regulations (e.g., ATT), we contribute to the growing body of studies on privacy, especially work focused on privacy regulations such as the GDPR and ATT. To date, most prior studies have examined the direct effects of data regulations on firms, including outcomes such as web traffic (Goldberg et al., 2019), firm revenue (Aridor et al., 2020), innovation and venture investment (Bessen et al., 2020; Janssen et al., 2022; Jia et al., 2021), SDK usage in Android mobile apps (Jin et al., 2024), data reliance (Demirer et al., 2024), and firms' choice of web technology (Peukert et al., 2022). Research focusing on the impact of ATT and third-party cookie ban has examined how ATT alters firms' monetization choices (Kesler, 2023; Aridor et al., 2024), advertising effectiveness (Alcobendas et al., 2023; Aridor et al., 2024; Wernerfelt et al., 2024), app market concentration (Li and Tsai, 2022), consumer opt-in rates (Kraft et al., 2023), and financial fraud resulting from excessive data sharing (Bian et al., 2023).[8]

Few studies examine spillover effects in the data economy. Notably, Aridor et al. (2020) documents consumer-side spillovers, showing that consumer privacy decisions enable firms to infer other consumers' types. Using ATT as a shock to the strength of inter-firm data linkages, our paper complements this work by documenting firm-side data externalities, demonstrating that firms' data collection and product design choices affect other firms within the data-sharing network.

Our paper also relates to the burgeoning literature on cyberattacks. Crosignani et al. (2023) document the propagation of cyber attacks through firms' supply chains. Akey et al. (2023) examines the impact of cyber events on firm value. Our paper adds to this literature by documenting the propagation of the negative impact of cyberattacks through the data network. To this end, our

[7]Our metric based on the data-sharing network contributes to the broader literature on measuring systemic risk (Billio et al., 2012; Acharya et al., 2016; Adrian and Brunnermeier, 2016; Benoit et al., 2016; Bai et al., 2018; Duarte and Eisenbach, 2021; Greenwood et al., 2015).

[8]See Johnson (2022) for a review of the literature on GDPR.

paper also relates to the broader literature on transmission of different shocks in inter-firm production networks, including productivity disturbances (Boehm et al., 2019; Barrot and Sauvagnat, 2016; Carvalho et al., 2021), financial shocks (Demir et al., 2024), and monetary policy and inflation shocks (Auer et al., 2019; Ozdagli and Weber, 2017). We study a new form of shocks and characterize its transmission in the novel and increasingly important data-driven networks.

2 Data Sharing Network: Measurement and Evidence

2.1 Institutional background

Data economy and mobile applications. Digital economy has emerged as a key pillar of the broader U.S. economy. In 2022, the Bureau of Economic Analysis (BEA) estimated its value at nearly $2.6 trillion, with an annual growth rate of 7.1% since 2017. It now accounts for 10.0% of U.S. GDP and supports 8.9 million jobs.[9] The rise of the digital economy is enabled by the increasing usage of mobile devices, a trend accelerated during the COVID-19 pandemic when demand for digital services across work, entertainment, and communication skyrocketed. After the pandemic in 2022, the average time spent by U.S. adults on mobile devices rose 2.5% year-over-year to over 4.5 hours per day, compared to just 3 hours and 7 minutes per day spent watching traditional TV.[10] This shift reflects a structural change toward a more digital-centric lifestyle.

Data has become an essential productive asset in the digital economy. Companies rely on mobile data to understand consumer preferences, customize product offerings, and guide innovation choices. Bian et al. (2021) show that over 60% of apps tracks users across websites, apps, and offline stores. Binns et al. (2018) report that nearly 90% of Android apps collect user data and en-

[9]Source: https://www.bea.gov/data/special-topics/digital-economy
[10]Source: https://www.emarketer.com/content/us-time-spent-with-connected-devices-2022

able data-sharing with Google. Companies leveraging personal data for targeted advertising—such as Google and Meta—generated approximately $780 billion in revenue in 2023.[11]

A key feature of mobile-collected data is its high connectivity, facilitated by user identifiers like the Identifier for Advertisers (IDFA) assigned by Apple. IDFA enables app developers, advertisers, data analytics platforms, and ad networks to track user behavior across iOS apps, providing a consistent device-level identifier that simplified user tracking and data sharing.[12] In contrast, web-based data relies on fragmented cookie systems managed by individual websites and advertising networks, which are often blocked by browsers like Safari. Therefore, IDFA offers greater consistency for user tracking, making mobile data a increasingly valuable asset for firms.

Firms in the data economy. Driven by the growing importance of mobile apps in data collection and utilization, we define firms participating in the data economy as those owning at least one mobile app. Our study centers on U.S.-listed public companies with reliable accounting and stock performance data, identifying a total of 1,031 firms meeting these criteria.

Firms in the data economy exhibit distinct characteristics from the average firm in the Compustat universe. Figure 1 presents the representativeness of these two sets of firms across and within industries. Panel A displays the share of data-reliant firms within each of the Fama-French 48 industries, with red bars representing equal-weighted shares and blue bars representing size-weighted shares. Focusing on the size-weighted shares, data-reliant firms account for more than 50% of the total assets in 24 out of 48 industries, and over 60% of the total assets across all Compustat firms. Panel B further examines the distribution within the data-reliant firm sample, revealing an over-representation in industries such as Business Services, Retail, and Telecommunications, and an under-representation in Banks, Finance, and Oil industries, compared to the broader Com-

[11] Source: https://www.marknteladvisors.com/research-library/digital-marketing-market.html

[12] The Android counterpart is the Google Advertising ID (GAID), also called the Android Advertising ID (AAID).

pustat universe. Additionally, Panel A of Figure 2 shows that, on average, data-reliant firms are larger. Panel B further highlights that these firms utilize their assets more efficiently to generate sales, as indicated by their higher asset turnover ratios.

App tracking transparency (ATT). Apple's App Tracking Transparency (ATT) policy, introduced in April 2021, requires app developers to obtain explicit user consent before tracking user activity across different apps and websites. This policy has directly impacted how firms collect, share, and use data, particularly within the digital advertising ecosystem, where cross-app tracking has been essential for delivering personalized ads and optimizing marketing strategies.

Before ATT, firms commonly used software development kits (SDKs)—pre-built software components integrated into mobile applications that provide various functionalities, including data collection and sharing—often without needing explicit user consent. ATT disrupts these practices by restricting the flow of user-level data collected via SDKs. Technically, the policy requires apps to display a prompt asking for permission to track user activity through the IDFA that was described above. If the user declines tracking, SDKs are blocked from accessing the IDFA, effectively disabling across-app user profiling. Without this cross-app identifier, SDKs must rely on more fragmented, often anonymized data. This reduced ability to track users limits firms' capacity to optimize advertising efficiency and product customization.

Data sharing network. By aggregating data from different firms and facilitating customer profiling, SDKs become the de facto data-sharing platforms. Data analytics services offered by the SDKs facilitate firms' marketing, product design, and other areas of decision making. Using data contributed by firms, the SDKs can offer predictions, for example, on new customers' preferences for firms that seek to expand production. The SDKs typically do not charge fees. Firms essentially

pay with their data contributions, making the exchange inherently barter-like.[13]

Two firms that share data with the same set of SDKs are thus connected as their signals on customers are based on a common set of data that both firms contribute and is aggregated and analyzed by the SDKs. ATT weakens the interconnectedness among firms under data sharing. In Section 2.3, we construct measures of data interconnectedness for each pair of firms and characterize a network structure of the data economy. Our measure is based on firms' overlap in their reliance on data-related SDKs. We will show that such data connectedness generates comovements across various metrics of firm performances and that ATT reduces performance comovements.

An example of data sharing. We use Amazon and General Motors (GM) to illustrate how data linkages between the two firms may benefit one another. According to our measure introduced in Section 2, GM has the highest data linkage with Amazon among all firms, with both companies using the data-related SDKs offered by Google. GM collects extensive data on consumer behavior through vehicle purchases, financing, and connected-car technologies. By sharing user data with Google Ads (SDK), GM enables Google to refine ad targeting for Amazon.[14] For instance, GM's data on consumers car purchasing patterns helps Amazon focus ads on individuals likely to buy complementary products, such as home EV chargers, smart car accessories, or high-tech gadgets. This precision targeting enhances Amazon's ability to convert high-value customers while improving its inventory planning for automotive-related products, boosting overall sales and profitability.

In return, Amazon shares e-commerce data with Google, which GM leverages to better understand consumer preferences and market trends. As disclosed in the privacy labels of Amazon's iOS mobile app, Amazon collects user-generated product interactions, linked to user identities,

[13]By offering data aggregation and analytics services, SDKs stimulate firms' productive activities including advertisement. SDKs' main source of revenues come from taking a share of advertisement publishers' profits. Therefore, the SDKs' data analytics services, by fueling the whole advertisement ecosystem, allow the SDKs to make profits.

[14]The GMC app on Apple's App Store discloses in its privacy labels that it collects and shares location data, contact information, and user and device identifiers with third parties.

for third-party advertising. Insights into consumer purchasing trends, such as rising demand for EV-related products among specific demographic groups, help GM tailor its marketing strategies. This data also informs GM's product development and financing offers, in alignment with current consumer trends.

2.2 Data sources

Apptopia. Apptopia is an alternative data provider which collects, structures, and models data about the mobile apps market. We use two products from Apptopia. First, we acquire information about mobile app characteristics (e.g., category, age) and performance (e.g., downloads, active users, sessions length). Second, we use their SDK intelligence product, which provides data on the installation and removal dates of SDKs for each mobile app. An SDK provides tools and libraries to integrate specific features or services, such as data analytics or payment systems, into mobile or web applications. It facilitates app development by offering pre-built components and resources. Apptopia tracks the history of third-party SDK integrations and removals in mobile apps by analyzing publicly available app installation package. When a new version of an app is released, Apptopia reanalyzes the updated package to maintain an accurate and current record of the app's SDK profile.

Based on the SDK categories defined by Apptopia, we identify the four categories as related to data sharing: advertising networks, analytics, mobile marketing, and monetization categories. The majority of SDKs belong to the first two categories. SDKs in the *advertising network* category connect apps with advertisers to display ads. SDKs in the *analytics* category offer insights into user behavior, app performance, and engagement metrics, facilitating data-driven performance optimization. SDKs in the *mobile marketing* category assist in executing marketing campaigns by providing tools for user segmentation, messaging, and engagement to boost user acquisition

and retention. SDKs in the *monetization* category provide revenue generation options, such as subscriptions, in-app purchases, and rewards, often supplementing advertising networks.

We verify the accuracy of each SDK's functionality by reviewing its documentation and technical details available through GitHub repositories. To minimize computational complexity, we focus on the 50 most popular data-related SDKs, ranked by worldwide net installations as of 2021/04/26. The most popular data-related SDK accumulated 262,209 net installations (i.e., installations minus uninstallations), while the 50th most popular SDK accumulated only 4,369 net installations in the same period, indicating a high level of concentration. Restricting the list to the top 20 SDKs yields a similarly distributed measure of data connectedness, which we introduce later. In addition to this main measure, we also use app category and user overlap between app pairs to construct measures in firm overlap in the app product markets.

Text-based Network Industry Classifications (TNIC). Horizontal and vertical industry linkages are obtained from the Hoberg-Phillips Data Library (Hoberg and Phillips, 2010, 2016; Frésard et al., 2020). These measures are available at annual frequency.

Factset Revere. Supply chain relationships are constructed from FactSet Revere – Supply Chain Relationships datasets. For each firm, we observe four types of relationships: customer, supplier, partner, and competitor. We also construct geographical overlap using the FactSet Revere – Geographic Revenue Exposure datasets, which offers revenue breakdown by geography and business segment. Both sets of measures are created at an annual frequency.

USPTO. We construct firms' technology proximity using data on their patent applications from the USPTO. For each firm's patent portfolio, we calculate technology proximity following the methodology of Jaffe (1986) and Bloom et al. (2013). This measure varies at annual frequency.

Standard financial datasets. Firms with shared financial analysts may exhibit performance co-movement (Ali and Hirshleifer, 2020). We obtain data on analyst coverage from I/B/E/S. Quarterly accounting data on firm fundamentals is from Compustat and data on stock prices and market capitalization from CRSP. Information on asset pricing factors is from Ken French's data library.

2.3 Variable construction

Data connectedness. To characterize the data-sharing network, we develop a firm-pairwise measure for data connectedness. This methodology is akin to the approach taken by Hoberg and Phillips (2010, 2016), which develop a product space overlap measure to study competitive interactions among firms; likewise, Bloom et al. (2013) introduce an R&D space overlap measure to explore the impact of technological proximity on innovation and firm performance. In our case, we compute the cosine similarity for pairs of firms based on their usage of data-related SDKs.

Specifically, we denote app a's (from firm i) data collected and shared with any SDK k at time (quarter) t as: $s_{iakt} = m_{iat} \times d_{iakt}$, where m_{iat} is number of monthly active users (averaged within quarter t) of firm i's app a; d_{iakt} = 1 if this app installs SDK k at time t and 0 otherwise. Aggregating across all apps owned by firm i at time t, we define the data shared with SDK k at firm-SDK-quarter level as: $S_{ikt} = \sum_{a \in \mathcal{A}_{it}} s_{iakt}$ where $\mathcal{A}_{it}$ is the set of apps owned by firm i at time t. We then stack the firm i's relationship with each relevant SDK at time t into a $K \times 1$ vector: $\mathbf{S_{it}} = \left[S_{i1t}, S_{i2t}, ..., S_{iKt}\right]'$, where where K is the total number of SDKs that serve as data-sharing intermediaries. K is set to 50 in the baseline version. The cosines-similarity between firm i and j in the data space at time t is given by:

$$\rho_{ijt}^{data} = \frac{\mathbf{S'_{it}} \cdot \mathbf{S_{jt}}}{|\mathbf{S_{it}}| \cdot |\mathbf{S_{jt}}|}, \tag{1}$$

where $|\cdot|$ is the Euclidean distance.

We further illustrate the construction of the data connectedness measure using Amazon as an example. Figure 3 displays Amazon and the 10 firms most closely connected to it through data. Firms are represented as blue circles, with the circle size corresponding to each firm's average monthly active users (MAU). Data-related SDKs are shown as red or yellow rectangles, and the thickness of the line connecting each firm to an SDK represents the SDK's relative importance to the firm, measured by the proportion of the firm's MAU linked to the SDK. Each of these 10 firms has a ρ^{data} of approximately 0.6 with Amazon, driven by the share of MAUs connected through two key SDKs – Firebase and Answers – that are commonly installed by both Amazon and these firms. However, their data connectedness with Amazon is not perfect, as each firm has also installed other data-related SDKs that Amazon has not installed (or vice versa), shown in yellow on the graph.

Notably, the firms highly connected to Amazon come from a diverse set of industries, including Destination XL, Vipshop, and Hibbett from retail; American Express from business services; Delta Air Lines from transportation; Mimecast and Vodafone from telecommunications; Morgan Stanley from banking; General Motors from automobiles and trucks; and Strategic Education from personal services. This variety supports our assertion that the data connectedness measure captures a novel, unique form of firm linkages, distinct from traditional connections within an industry or along the supply chain. In our empirical analysis, we control for other forms of interconnectedness.

Other firm linkages. We measure the proximity between two firms across various dimensions, including product markets, supply chains, innovations, analyst coverage, and geography. While app user overlap, supply chain relationships and analyst coverage are discrete variables, taking values of 0 or 1, the remaining proximity measures are continuous variables ranging from 0 to 1.

Performance comovement. We measure the performance comovement between two firms by calculating the correlation across various metrics, including the logarithm of downloads and daily active users (DAU), quarterly earnings growth, and asset turnover (sales/assets), all computed at a quarterly frequency.[15] For each performance metric, we compute one correlation for quarters before the implementation of ATT in 2021Q2 and another for the periods after the implementation.

We also consider comovement in firm stock returns. For each firm pair, return comovement is measured as the correlation of the pair's monthly returns in a rolling 12-month window from 2014 to 2024.[16] There are ten 12-month non-overlapping windows, 7 windows before the introduction of ATT in April 2021 and 3 windows afterwards. Therefore, for each pair, we have ten values of return comovement. We consider three types of returns: raw returns, abnormal returns based on CAPM, and DGTW-adjusted returns (Daniel et al., 1997).

Non-data SDK usage. Lastly, we construct firm-level variables related to the usage of SDKs that represent firms' endogenous product design decisions, ranging from monetization to customer engagement and data accumulation, as outlined in the model in section 4. For each of the following SDK categories – payment, security, and customer support – we calculate: 1) the number of unique SDKs used by a firm, 2) the change in the number of unique SDKs used by a firm, and 3) the weighted sum of the number of unique SDKs used by peer firms that are data-connected to the focal firm, where the weight is the pairwise data connectedness described above. The third measure captures the average product-design decisions made by the focal firm's data-connected peers.

[15] Quarterly earnings growth rate is calculated as $2 * (\textit{Net Income}_t - \textit{Net Income}_t)/(\textit{Net Income}_t + \textit{Net Income}_t)$, instead of $(\textit{Net Income}_t - \textit{Net Income}_t)/\textit{Net Income}_t$ to smooth out volatility.

[16] For example, the three correlations after ATT correspond to the 12-month periods from April 2021 to March 2022, April 2022 to March 2023, and April 2023 to March 2024.

2.4 Descriptive statistics

Data-sharing network. In Figure 4, we present the network structure implied by the data connectedness measure. This figure visualizes the network of firms connected through data sharing using the Fruchterman-Reingold Algorithm. The pairwise data connectedness takes the average value in 2020.[17] Each node represents a firm, with the firm's ticker displayed, and the size of the node corresponds to the firm's size, proxied by the square root of total MAU (monthly average users) in 2020. Firms linked by edges are those with strictly positive data connectedness. For readability, we only include firm pairs with data connectedness greater than 0.7, which includes 640 unique firms (72% of all firms active in the data network) and around 6.2% of all firm pairs. We label the firms that have more than 4.7 million MAU, or the 92^{th} percentile of the MAU distribution, in an average quarter in 2020. Firms situated at the center typically have more highly-connected peers. Firms are clustered into different colors as determined by the Louvain Community Detection Algorithm. We identify and label the most popular SDK within each cluster of firms.

Google (ticker: GOOGL) and Meta (ticker: META) appear to be the two most influential firms in the data network, evident from both their node size and central location. Consistent with our observation that data-reliant firms are overrepresented within the business services industry, many of these firms are also located near the network center, including Yandex (ticker: YNDX), Twitter (ticker: TWTR), Unity (ticker: U), and Cheetah Mobile (ticker: CMCM). Also positioned at the center of the network are firms from other industries, such as AT&T (ticker: T) from telecoms, and Electronic Arts (ticker: EA) and Roku (ticker: ROKU) from entertainment.

It is important to note that a firm's size does not always correlate with its network centrality. For example, large firms like Walmart, Nike, and Netflix remain on the periphery, while many

[17] Based on Spearman or Pearson correlation measures, the data network's time-series stability is comparable to that of the Hoberg–Phillip vertical or horizontal product network. In the empirical analysis, we use the average pre-ATT data connectedness and treat the network as fixed.

relatively smaller firms occupy central positions in the network due to their interconnectedness. Other relatively large firms, such as Snap, Pinterest, Zoom, and Apple, are excluded from this graph because their data connectedness with other firms is below 0.7.

While these patterns are intuitive, centrality visualized in this graph is only based on the direct linkages. Data sharing induces interconnectedness of higher orders—firm A's data is shared through SDKs with firm B that in turn may share data with firm C. Moreover, data spillover has persistent impact over multiple time horizons as data affects firms' ability to stimulate customer engagement which in turn contributes to further data accumulation resulting in self-perpetuating dynamics. In Section 4, we develop a framework to identify systemically important firms in the data economy that accounts for indirect linkages and data spillover effects across multiple horizons.

Other statistics Table 1 provides summary statistics for the variables in our regression sample, which includes 1,031 firms over the period from 2014Q3 to 2023Q2. Panel A lists all firm linkages. On average, firm pairs have a data similarity score ("data connectedness") of 0.170, with the 90^{th} percentile at 0.46. The app-category similarity and similarity in geographical distribution of business segments also exhibit relatively large means of 0.157 and 0.300, respectively. Additionally, 0.8% of firm pairs have customer-supplier linkages, and 6.3% of firms share common analysts.

The correlation in app and financial performances varies substantially across firm pairs. For example, the 10^{th} and 90^{th} percentile of correlation in log(downloads) are -0.67 and 0.73, respectively. We also report the correlations for stock returns, CAPM-adjusted returns, and DGTW-adjusted returns. There is stronger comovement in raw returns (0.248) compared to abnormal returns based on CAPM (0.034) or DGTW-adjusted returns (0.008). This is because raw returns capture exposure to common risk factors. All three measures of return comovement exhibit significant variations across firm pairs, with the 90^{th} percentile at 0.67, 0.47, and 0.44, respectively.[18]

[18]The pairwise observations of return correlations are more than those of firm performance correlations. For return

Finally, in Panel B of Table 1, we report the summary statistics for firm-level variables, including changes in a firm's SDK usage and the stock of peer firms' SDKs, categorized by the functionality of SDKs. The average firm has 2.023 apps actively using payment SDKs and experiences a change of 0.008 in the number of apps with active payment SDKs. Additionally, the average firm is connected to peer firms that have 4.391 apps actively using payment SDKs. This set of measures are motivated by our model in section 4. In terms of firm characteristics, the average firm in our sample has a long term debt ratio of 26.2%, a tangibility of 20.3%, and a cash-to-asset ratio of 18.6%. In comparison, the average firm in the non-app sample has a long term debt ratio of 27.3%, a tangibility of 26.1%, and a cash-to-asset ratio of 10.3%.

3 Comovement and Shock Propagation under Data Sharing

The customer-profiling signals that firms receive from SDKs contain data from one another. Since data is often generated as a by-product of business operations (e.g., Bergemann and Bonatti, 2019; Jones and Tonetti, 2020; Farboodi and Veldkamp, 2021), the expansion of one firm brings more data to its SDK-connected peers. Being more informed about customers allows the peer firms to expand operations and improve profitability. In this section, we examine performance comovements induced by data sharing, and using cyberattacks as our empirical setting, we trace out how shocks are propagated through the data-sharing network, generating ripple effects across firms.

correlations, we have ten observations for each pair as previously described—we compute correlation of monthly returns for each of ten 12-month windows—while for performance correlation, we compute one value for quarterly observations in the pre-ATT window and one value for the post-ATT window (two values per pair).

3.1 Performance comovement

Data allows firms to improve performances in various areas from optimizing advertisement to customizing product offerings. We test the hypothesis that two firms that share data with one another through the SDKs are likely to exhibit comovement in their operational and financial performances. In addition, the introduction of ATT disrupts data sharing and thereby weakens such performance comovement. Therefore, we also examine the impact of ATT. Our empirical specification follows the gravity model in the international trade literature (e.g., Imbs, 2004; Baxter and Kouparitsas, 2005; Di Giovanni and Levchenko, 2010). Specifically, we estimate the following regression:

$$Corr^P_{ijt} = \alpha + \beta_1 \rho^{data}_{ij} + \beta_2 ATT_t \times \rho^{data}_{ij} + \beta_3 \rho^{other}_{ij} + \beta_4 ATT_t \times \rho^{other}_{ij} + \theta_{it} + \iota_{jt} + \varepsilon_{ijt}\,. \quad (2)$$

The correlation metrics, $Corr^P$, capture firm pairwise comovement across multiple dimensions, where P is app performance, financial performance, or stock returns. For each pair ij and comovement in app and financial performance, we have two observations, one before ATT and one after, with $t \in \{\text{pre-ATT, post-ATT}\}$. For return comovement, we have one observation per 12-month window and a total of ten observations (seven before ATT and three after ATT). We include firm-by-time fixed effects (θ_{it} and ι_{jt}) and double cluster standard errors by firm-i and firm-j.

To ensure that our results are not confounded by other forms of linkages between firms, we include controls for linkages, denoted as ρ^{other}_{ij}, that have been shown in prior literature to impact comovement. Specifically, we control for product market overlap, supply chain relationships, technological proximity, and common analyst coverage, among others. For all the firm linkages, including our measure of data connectedness, ρ^{data}_{ij}, we use the average value before ATT, as the post-ATT values may reflect firms adjusting data connections in response to ATT.

The coefficients of interest are β_1 and β_2. A positive value of β_1 indicates that a high degree

of data connectedness (i.e., a high ρ_{ij}^{data}) is associated with more synchronized variations in firms' performances. Moreover, we expect β_2 to be negative, as ATT restricts data flows between firms, thereby weakening the performance comovements between firms. To facilitate the interpretation of the coefficients and comparison across different types of firm linkages, we normalize ρ_{ij}^{data} and ρ_{ij}^{other} to have a zero mean and a standard deviation equal to one.

The results are reported in Table 2 to Table 4. Starting with app performance comovement, Table 2 confirms that higher data connectedness is associated with a significantly greater degree of app performance comovement. For example, based on column 1 of Table 2, a one-standard-deviation increase in data connectedness leads to a 0.025-unit increase in the correlation of log(downloads) between two firms. Importantly, this relationship disappears after the implementation of ATT, with the coefficient of $ATT \times \rho_{ij}^{data}$ at -0.025, offsetting the baseline effect. Including other types of firm linkages and their interaction terms with ATT has little impact on the estimation results. We find similar results when focusing on the comovement in log(DAU).

Because the distributions of firms' pairwise correlations tend to be centered around zero, directly comparing the coefficient to the average would not be informative. Instead, we interpret the magnitude of these effects by comparing them to other firm linkages. For example, based on Column 2, the estimated coefficient of 0.024 for ρ_{ij}^{data} is 2.6 times larger than that for firms' horizontal overlap in product markets based on the TNIC, which is 0.009. It is noteworthy that the coefficients for the interaction terms between ATT and ρ_{ij}^{other} are almost never negative or economically significant, with the exception of $ATT \times \rho_{ij}^{product\,horizontal}$ in column 4. This supports our interpretation that ATT primarily affects firm performance by limiting data sharing.

Turning to the comovement in financial performance in Table 3, we find that a one-standard-deviation increase in data connectedness is associated with a 0.002 increase in the comovement of earnings growth before ATT. However, this effect is completely reversed after ATT, with the coef-

ficient of $ATT \times \rho_{ij}^{data}$ at -0.003. The magnitudes are similar when examining the comovement in asset turnover. Based on column 2, the effect of data connectedness on performance comovement is approximately 40% of that of horizontal overlap in product markets, as measured by the TNIC. The smaller effect of data connectedness on financial performance compared to app performance is intuitive, as firms may have other non-app business segments.

Finally, we examine stock return comovement. Similar to previous tables, Table 4 reports the results for raw returns (columns 1-2), abnormal returns based on CAPM (columns 3-4), and DGTW-adjusted returns (columns 5-6), with and without other types of firm linkages as control variables. The results are largely consistent across the three types of returns. For instance, based on the DGTW-adjusted returns in column 6, a one-standard-deviation increase in data connectedness is associated with a 0.002 increase in return comovement, which is about 12% of that of horizontal product market overlap. After ATT, this effect of data connectedness on return comovement is reduced significantly. Notably, none of the alternative linkages experience a sharp drop in their effects on return comovement after ATT, except for mobile-user overlap in column 4, which captures firms' similarity in the app product space and therefore can be correlated with data connectedness.

3.2 Cyberattack ripple effects

When a firm is hit by a shock that affects its operations and data collection, the impact is transmitted to its data-connected peers, generating performance comovement between firms. In other words, the origin of comovement is shock propagation. Next, we examine how the data-sharing linkages propagate the impact of cyberattack, a type of shocks that is particularly relevant for data economy.

We identify major cyberattack events using data from Advisen. This dataset covers more than 90,000 cyber events between 2000 and 2023 collected from publicly verifiable sources, including

government websites, keyword-based searches, and official court and litigation sources.[19] We identify 22 major cyber events that involve at least 10 million exposed records. The list of these events and their summaries can be found in Table B.1 in the internet appendix.

For each firm k involved in a cyber event, we define a peer firm i's exposure to the event as:

$$\text{Exposure}_{ik} = \frac{\sum_P \rho_{ik}^{\text{data},P} DAU_k^P}{\sum_P \sum_j \rho_{ij}^{\text{data},P} DAU_j^P}$$

where j represents a firm connected to firm i via data sharing, and P represents platforms, taking values from {iOS, Android}, and $\rho_{ik}^{\text{data},P}$ is data connectedness between firm i and firm k, as defined in equation (1). A higher value of Exposure_{ik} implies that firm k's data is important relative to all other firms connected to firm i. A firm j is considered to be highly exposed to firm k's cyber event if $\text{Exposure}_{ik} \geqslant 0.01$, corresponding to the 75th percentile of the exposure distribution. Firms with $\text{Exposure}_{ik} < 0.01$ are considered control firms. We assign an indicator variable, *high exposure*$_{ik}$, which takes the value of 1 for highly exposed firms and 0 otherwise.

We estimate the following regression equation on the 22 major cyber events using a stacked difference-in-differences approach:

$$Y_{ikt} = \alpha + \beta_1 cyber\ event_k \times high\ exposure_{ik} + \beta_2 cyber\ event_k \times \rho_{ik}^{other} + \mathbf{X_{it}}\beta_3 + \theta_{ik} + \iota_{kt} + \varepsilon_{ikt}, \tag{3}$$

where k indexes events, i indexes firms, and t indexes the quarter relative to each event. For each event, we use an 16-quarter window, with 8 quarters before and 8 quarters after the event. We consider two outcomes for Y_{it}: total downloads and DAU (daily active users) for a firm in a given quarter, both in natural logarithm. cyber event$_k$ is an indicator variable that takes the value of 1 after the cyber event occurs. ρ_{ik}^{other} includes non-data firm linkages between i and k. In all

[19] More information on Advisen's data sources can be found on their website.

specifications, we include the following firm-level controls, $\mathbf{X_{it}}$: firm size (log of assets), long-term debt to assets, and tangible assets to total assets. Additionally, we control for firm$\times$event fixed effects θ_{ik} and event-specific relative quarter fixed effects ι_{kt}. The firm$\times$event fixed effects ensures that the identification of our coefficient of interest β_1 relies on comparing the treated and control firms within each event. Standard errors are double-clustered by event and firm.

We report the estimation results in Table 5, with the first two columns presenting the results on downloads and the last two columns on DAU. In all columns, β_1 is negative and statistically significant at the 1% level. The magnitude of the cross-firm spillover effect of major cyber events is economically large. Specifically, relative to the control firms, the highly exposed firms experience a 8.3% drop in quarterly downloads (column 1) and a 9.6% drop in DAU (column 3). Controlling for the interaction of cyber event$_k$ with other firm linkages, these point estimates drop slightly to 6.8% and 7.7%, as reported in column 2 and column 4, respectively. Additionally, when estimating the dynamic version of equation (3), we show in Figure 5 that our results are not driven by pre-trends between control and treated firms. There are no discernible pre-event trends, and the treatment effect begins to emerge immediately after the cyber event occurs, continuing through the fourth quarter post-event. Finally, it stabilizes after five quarters with a slight upward reversal.

Next, we present an event study on peer firms' stock performance surrounding the cyber events. The results are shown in Figure 6, where Panel A and Panel B use the incident dates and notice dates, respectively, as the event dates. Both panels illustrate the cumulative abnormal returns for peer stocks with high exposure to the event, using CAPM as the benchmark model. Major cyber events and highly exposed peer firms are defined the same way as before. As shown in these figures, firms connected to the focal firms experience a 3-4% decline in cumulative abnormal returns in the month following either the incident or notice date. Using other benchmark models, such as the Fama-French three factor model and DGTW, yields qualitatively similar results.

These findings provide the first evidence of how negative shocks to firms spill over to their peers through the data-sharing network, hindering peers' operational performance and valuation.

4 A Network Model of Data Economy

Motivated by our empirical findings in Section 2, we develop a model of data economy where firms are interconnected through data sharing. Data plays two roles in our model. First, it enables firms to stimulate customer engagement. Second, it reduces firms' costs of targeting customers for monetization. Therefore, data, by revealing customer's preferences, allows a firm to interact more actively with customers and to generate profits through such interaction. Each firm accumulates data, and through data analytics platforms, shares data with other firms. ATT disrupts data sharing.

4.1 The setup

Data dynamics. The economy has N firms. We consider firm i's problem, $i \in \{1, ..., N\}$. Let $\delta_{i,t}$ denote the firm's stock of data on its customers. The firm collects data from interaction with customers ("customer activities"), denoted by $y_{i,t}$, and $\delta_{i,t}$ evolves as

$$d\delta_{i,t} = \theta y_{i,t} dt + \mu_\delta \delta_{i,t} dt + \sigma_{i,\delta} \delta_{i,t} dz_{i,t}, \tag{4}$$

where the last two terms reflect a stochastic growth rate of data that is independent from the current customer activities, and $z_{i,t}$ is a standard Brownian motion that is independent across firms.

The first term on the right side of equation (4), $y_{i,t}dt$, captures the idea that data is a by-product of customer activities, where θ (> 0) represents data generation efficiency (e.g., Bergemann and Bonatti, 2019; Farboodi and Veldkamp, 2021). For example, an e-commerce platform

collecting data on consumers by observing their transactions and search activities. And, for a retail manufacturer, the marketing and sales efforts generate interactions with customers, and feedback from customers is informative of their preferences. Additionally, software companies learn about users' preferences through their usage patterns. Our empirical setting covers firms across industries. Accordingly, we set up our model to be sufficiently generic to highlight the commonality among these firms, that is their reliance on data and sharing data with one another.

The parameter, μ_δ, can be negative in which case data depreciate. The firm uses data to profile customers. A higher customer turnover rate and a more volatile customer preferences are likely to be associated with faster data depreciation (i.e., a lower μ_δ), as stale data becomes less informative. Finally, the diffusion term, $\sigma_{i,\delta} dz_{i,t}$, captures shocks to firm i's data stock. A negative shock may reflect a direct loss of data, for example, due to cyberattack or regulatory and legal actions against data use.[20] In contrast, following a positive shock, the firm gains more information on its customers, for example, through an increase in product reviews and improved access to technologies that enable data collection. Our model can be used for analyzing how shocks to one firm spills over to other firms through the data-sharing network that amplifies the aggregate impact.

Customer engagement. Next, we model how customer activities, $y_{i,t}$, is determined. In the data law of motion (4), $y_{i,t}$ contributes to data accumulation. Data in turn contributes positively to the firm's ability to stimulate customer activities as data allows the firm to be informed about its customers. For example, a retail manufacturer can interact with its customers and stimulate customer activities only if it knows where the customers shop, post reviews, and view advertisements in physical locations and on the internet. For software companies, knowing customers' preferences and habits is critical for increasing user engagement and time spent on the product.

The firm transmits its data through software development kits (SDKs) to data analytics plat-

[20] Florackis et al. (2022) measure cybersecurity risk from corporate disclosure.

forms and in return obtains a composite signal on its customers, which combines data from different firms. For firm i, we define $D_{i,t}$ as the composite signal for firm i:

$$D_{i,t} = \sum_{j=1}^{N} \gamma_{ij} \delta_{j,t}, \tag{5}$$

where $\sum_{j=1}^{N} \gamma_{ij} = 1$ with γ_{ii} representing the weight on firm i's own data and γ_{ij} representing the weight on firm j's data, $\delta_{j,t}$, $j \neq i$.

Next, we specify the technology for stimulating customer activities: $y_{i,t}$ is determined by

$$y_{i,t} = \alpha D_{i,t} + x_{i,t}, \tag{6}$$

where α is a parameter that links $D_{i,t}$, i.e., how informed the firm is about its customers, to customer engagement. The firm's own data, $\delta_{i,t}$, contributes to $y_{i,t}$ through the composite signal, $D_{i,t}$.

The signal obtained from data analytics platforms generates a baseline level of customer activities, $\alpha D_{i,t}$, and the firm can increase customer engagement by an extra amount, $x_{i,t}$, by adjusting its product design. When choosing a higher $x_{i,t}$, the firm makes its product more suitable for generating customer engagements, for example, by offering free services, rather than monetization; in other words, embedded in the choice of product design is a trade-off between monetization and customer engagement, which we will specify when discussing firm i's cash-flow generation.

Substituting (5) and (6) into (4), we summarize the data dynamics block of our model in the following equation for the growth rate of firm i's data stock:

$$\frac{d\delta_{i,t}}{\delta_{i,t}} = \left(\theta\alpha \sum_{j=1}^{N} \gamma_{ij} \frac{\delta_{j,t}}{\delta_{i,t}} + \theta \frac{x_{i,t}}{\delta_{i,t}} \right) dt + \mu_\delta dt + \sigma_{i,\delta} dz_{i,t}, \tag{7}$$

A small firm with a lower $\delta_{i,t}$ tends to benefit more from other relatively larger firms' data as the

ratio $\delta_{j,t}/\delta_{i,t}$ tends to be greater in the expected growth rate of $\delta_{i,t}$. Therefore, data sharing is an equalizing force in the economy. ATT reduces θ and weakens this force. Other privacy regulations (e.g., GDPR) have similar effects. In line with this model feature, empirical evidence suggests that such regulations contributes to market concentration (e.g., Johnson et al., 2023; Jia et al., 2021).

By reducing θ, ATT also weakens firms' incentive to increase $x_{i,t}$, i.e., to prioritize customer engagement and data accumulation over monetization. The intertemporal trade-off a firm faces is that by increasing $x_{i,t}$, a firm sacrifices monetization now but through a higher $\delta_{i,t}$ may be able to monetize more in the future. Next, we specify how firms' cash flows depend on the choice of $x_{i,t}$.

Cash flows. At time t, firm i has $\omega_{i,t}$ number of paying customers. Each customer contributes ζ dollars of profits.[21] The number of paying customers is in turn determined by two factors. The first factor is about the history of customer engagement and the second is the firm's marketing effort.

As shown in equation (4), $\delta_{i,t}$ represents firm i's data stock but also reflects the history of customer engagement by accumulating customer activities, $y_{i,t}$, over time. Let $\kappa\delta_{i,t}$ denote the number of paying customers generated by such customer capital, i.e., the first component of $\omega_{i,t}$. The associated profits are $\zeta\kappa\delta_{i,t}$. In summary, the parameter κ scales $\delta_{i,t}$ to the number of customers, and the parameter ζ scales the number of customers to profit.

Here we can already see the interconnectedness from data sharing having an impact on firms' financial and operational performances and generates comovement. As shown in (4), (5), and (6), customer engagement, $y_{i,t}$, increases and $\delta_{i,t}$ grows faster when firm i is more informed about its customers (i.e., $D_{i,t}$ is higher). The composite signal, $D_{i,t}$, in turn depends on other firms' $\delta_{j,t}$ through the data-sharing network where firm i's dependence on firm j's data is given by γ_{ij}.

Our model setup is motivated by the empirical findings in Section 2 and captures the comove-

[21] In Appendix C.1, we provide a simple microfoundation based customers' optimal choices of spending on the firm's product and the firm's optimal pricing decisions.

ment in firms' performances that is driven by data sharing. As shown in Section 2, ATT reduces comovement from data sharing. In our model, ATT maps to a reduction of θ, the parameter that governs the amount of data a firm can collect from customer activities. A lower θ weakens firms' interconnectedness under data sharing: other firms' data affects firm i through $D_{i,t}$, which is an input into firm i's customer activities, $y_{i,t}$, but under a lower θ, customer activities, $y_{i,t}$, contributes less to firm i's data growth (hence other firms' data contributes less to firm i's data growth).

We have discussed the role of customer capital (cumulative customer engagement) in generating profits. Next, we introduce the second factor—marketing efforts—that contributes to the acquisition of paying customers. Let $\omega_{i,t}^m$ denote the number of paying customers attributed to marketing efforts (hence the total customer base is $\kappa\delta_{i,t} + \omega_{i,t}^m$). Let $e_{i,t}$ denote the firm's market efforts, and we assume that $\omega_{i,t}^m = \omega^m(e_{i,t})$ is increasing and concave in $e_{i,t}$.

The firm chooses $e_{i,t}$ to maximize the profits net off marketing costs

$$F_{i,t} = \max_{e_{i,t}} \ \zeta\omega_{i,t}^m - C(D_{i,t}, x_{i,t})e_{i,t}, \tag{8}$$

where $F_{i,t}$ represents the net profits. The unit cost of marketing effort is $C(D_{i,t}, x_{i,t})$ with the following properties. First, $C_D < 0$ and $C_{DD} > 0$: the composite signal, $D_{i,t}$, reduces the effort cost but its marginal benefit is decreasing.[22] This is in line with the decreasing marginal benefit to data in forecasting precision in Farboodi and Veldkamp (2021).[23] Second, $C_x > 0$ and $C_{xx} > 0$: the effort cost is increasing and convex in $x_{i,t}$. Given a product design that prioritizes customer engagement over monetization (i.e., a high $x_{i,t}$), it is difficult for the firm to induce customers to pay. For example, when a majority of functionalities are offered free, a software company must devote significant efforts to induce customers to pay for the very few premium features.

[22]This is consistent with the literature that models intangibles as a factor of production that enables firms to lower the cost of entering new markets (Argente et al., 2021; Hsieh and Rossi-Hansberg, 2023).

[23]Data reveals individual customers' preferences but the potential heterogeneity across customers is finite.

Importantly, the cross-derivative of the cost function is negative ($C_{xD} < 0$), that is when the firm is more informed about its customers (i.e., $D_{i,t}$ is higher), the marginal cost of marketing efforts is less affected by product design choices, $x_{i,t}$. For example, when a software firm understands its customers better, it can advertises the premium features more effectively (to targeted customers) even though its product design prioritizes customer engagement by offering many free services.

Lemma 1 (Data, Product Design, and Profits) *The optimized $\omega^m_{i,t}$ and associated profits, $F_{i,t}$, are increasing in $D_{i,t}$, i.e., $\frac{\partial \omega^m_{i,t}}{\partial D_{i,t}} > 0$ and $\frac{\partial F_{i,t}}{\partial D_{i,t}} > 0$, and decreasing in $x_{i,t}$, i.e., $\frac{\partial \omega^m_{i,t}}{\partial x_{i,t}} < 0$ and $\frac{\partial F_{i,t}}{\partial x_{i,t}} < 0$. Additionally, the cross derivative of $F_{i,t}$ is positive, i.e., $\frac{\partial^2 F_{i,t}}{\partial x_{i,t} \partial D_{i,t}} > 0$.*

In summary, the firm' profits are $\zeta\kappa\delta_{i,t} + F_{i,t}$, where the first component reflects profits from the classic channel of customer capital (built through past customer engagement that is proportional to $\delta_{i,t}$) and the second component, $F_{i,t}$, depends on the active marketing efforts. The composite signal $D_{i,t}$ from data analytics platforms contributes to $F_{i,t}$ by making the marketing efforts more targeted and efficient.[24] Designing the product to downplay monetization, i.e., increasing $x_{i,t}$, directly reduces $F_{i,t}$ by making it more difficult to acquire paying customers.[25] The positive cross derivative, $\frac{\partial^2 F_{i,t}}{\partial x_{i,t} \partial D_{i,t}} > 0$, suggests that when firm i is more informed about its customers, this negative impact of increasing $x_{i,t}$ on profits is mitigated. This property inherits from $C_{xD} < 0$.

Figure 7 illustrates the three blocks of our model, data dynamics, firm product-design and marketing decisions, and firm valuation that we solve in Section 4.2. The choice is $x_{i,t}$ is akin to an investment decision, involving an intertemporal trade-off: increasing $x_{i,t}$ reduces current profits

[24] While we model the role of data as cost reduction, improved profitability may also come from data-enabled price discrimination (Ichihashi, 2020). The associated harmful impact on customers is beyond the scope of this paper.

[25] Increasing $x_{i,t}$ to stimulate customer engagement and data accumulation at the expense of monetization is a form of "active experimentation" in line with Farboodi et al. (2019). In Farboodi et al. (2019), the impact of experimentation on profits can be positive or negative (rather than always negative as in our model), but the impact is negative at the optimum: the optimal scale of experimentation (the choice of production scale in their model) is always sufficiently large such that the marginal impact on current profits is negative. The firm is willing to accept the negative impact on profits because the marginal value of data acquired through experimentation is positive. In our model, we explicitly focus on the region where such an explicit trade-off between current profits and data acquisition emerges.

but enhances customer engagement, $y_{i,t}$, and the accumulation of data, $\delta_{i,t}$ (see equation (4)). A higher $\delta_{i,t}$ in the future boosts future profits, because both components of profits, $\zeta\kappa\delta_{i,t}$ and $F_{i,t}$, depend on $\delta_{i,t}$ via $D_{i,t}$ (see the definition (5)). The data stock, $\delta_{i,t}$, plays the role of "productive capital". If a firm is more informed (i.e., $D_{i,t}$ is higher because either its own data stock is higher or its connected peers have more data), it faces a lower marginal cost of investment under $C_{xD} < 0$.

In the firm-decision block, our focus is on $x_{i,t}$, i.e., the data investment decision, rather than $e_{i,t}$, the marketing efforts. Given $x_{i,t}$, the choice of $e_{i,t}$ in (8) simply leads to the profit function $F_{i,t}(D_{i,t}, x_{i,t})$ that has the properties in Lemma 1, where $\frac{\partial F_{i,t}}{\partial x_{i,t}} < 0$ reflects the intertemporal trade-off between current profits and data accumulation for future profit generation and, importantly, this tension is mitigated when the customer-profiling signal, $D_{i,t}$, is higher, i.e., $\frac{\partial^2 F_{i,t}}{\partial x_{i,t}\partial D_{i,t}} > 0$.

Our model differs from the classic investment theories (e.g., Hayashi, 1982; Abel and Eberly, 1994) in two key aspects. First, data investment has a positive externality.[26] Second, firms' investment decisions are interconnected through data sharing. In the next subsection, we show that a firm's data marginal q (the derivative of value function with respect to data stock) drives the optimal $x_{i,t}$ and incorporates the expected trajectories of data inflows from other firms. Let ρ denote the discount rate. We define the firm's valuation (i.e., the value function at $t = 0$):

$$\rho V^i\left(\delta_{i,0}, \{\delta_{j,0}\}_{j\neq i}\right) = \max_{\{x_{i,t}, e_{i,t}\}_{t=0}^{\infty}} \mathbb{E}\int_{t=0}^{+\infty} e^{-\rho t}(\zeta\kappa\delta_{i,t} + F_{i,t})dt. \tag{9}$$

The state variables include firm i's data and other firms' data, which affects the composite signal.

Discussion: The characteristics of data assets. Our model captures the three features of data as a productive asset. First, data accumulates through customer activities and thus is a by-product of

[26]This stands in contrast to the traditional investment dynamics where one firm's investment often crowds out others' investment. For example, investment by one firm may increase the cost of investment inputs and cost of financing or intensifying product-market competition that other firms face (e.g., Asriyan et al., 2024).

firms' operations (e.g., Bergemann and Bonatti, 2019; Farboodi and Veldkamp, 2021).[27] Second, data is non-rival: sharing data with other firms does not prevent a firm from using its data or cause it to lose data (e.g., Jones and Tonetti, 2020). Third, data has externality: data on one firm's customers can be informative about other firm's customers (e.g., Ichihashi, 2020).

The second and third features explain why firms are willing to share data. Firm-level data externality derives from externality at the individual levels—one person sharing data reveals other people's attributes. Choi et al. (2019) and Acemoglu et al. (2022) point out that data externality leads to excessive data sharing and collection. In our model, firms under-invest in data accumulation under positive data externalities (and thus monetize excessively). The difference lies in the fact that in Choi et al. (2019) and Acemoglu et al. (2022), it is the consumers who decide on sharing data, while in our model, firms set the speed of data accumulation via product-design choices.

The combination of these three features distinguish data from other intangible assets. One example is R&D, which also generates knowledge that is a non-rival asset and has positive spillover effects (i.e., the second and third features of data). However, investing in R&D is costly and can be separated from production. In contrast, data is generated as by-product of business operations. While firms may incur costs to stimulate customer activities and generate more data, the baseline level of data generation is free, and therefore, through data sharing, there exists a baseline positive externality of one firm's production and the associated data generation on other firms.

4.2 Equilibrium

Firm i's own data stock, $\delta_{i,t}$, is a state variable, and, due to the dependence on other firms' data via the composite signal, $D_{i,t}$, the other firms' $\delta_{j,t}$ ($j \neq i$) are also state variables for firm i. We

[27]The notion that data is a byproduct of economic activity was well established in the information economics literature (e.g., Veldkamp, 2005; Ordoñez, 2013; Fajgelbaum et al., 2017).

use $F\left(D_{i,t}, x_{i,t}\right)$, defined in Lemma 1, to denote the maximized profits under the optimal choice of marketing efforts, $e_{i,t}$, given any value of $x_{i,t}$. $F\left(D_{i,t}, x_{i,t}\right)$ satisfy the properties in Lemma 1. Next we analyze the optimal choice of $x_{i,t}$ through the following Hamilton-Jacobi-Bellman (HJB) equation for the value function of firm i at time t, denoted by $V^i\left(\delta_{i,t}, \{\delta_{j,t}\}_{j\neq i}\right)$:

$$\rho V^i\left(\delta_{i,t}, \{\delta_{j,t}\}_{j\neq i}\right) = \max_{x_{i,t}} \zeta\kappa\delta_{i,t} + F\left(D_{i,t}, x_{i,t}\right) + V^i_{\delta_{i,t}}\left[\theta(\alpha D_{i,t} + x_{i,t}) + \mu_\delta \delta_{i,t}\right] + \frac{1}{2} V^i_{\delta_{i,t}\delta_{i,t}} \delta^2_{i,t}\sigma^2_{i,\delta} + \sum_{j\neq i}\left[V^i_{\delta_{j,t}}\left[\theta(\alpha D_{j,t} + x_{j,t}) + \mu_{j,\delta}\delta_{j,t}\right] + \frac{1}{2} V^i_{\delta_{j,t}\delta_{j,t}} \delta^2_{j,t}\sigma^2_{j,\delta}\right]. \tag{10}$$

Note that to highlight the intertemporal linkage between $x_{i,t}$ and data accumulation, we substitute out $y_{i,t}$ in the drift of $\delta_{i,t}$ using (6), i.e., $y_{i,t} = \alpha D_{i,t} + x_{i,t}$. The following proposition characterizes the optimal $x_{i,t}$ through a Q-theory of data investment. As previously discussed, the intertemporal trade-off is between the negative impact on the current profits, $F_x(D_{i,t}, x_{i,t})$, and the positive impact on data accumulation evaluated at the marginal value of data (or "*data marginal q*"), $V_{\delta_{i,t}}$.

Proposition 1 (Q-theory of Data Accumulation) *The first-order condition for $x_{i,t}$ is given by*

$$-F_x(D_{i,t}, x_{i,t}) = V^i_{\delta_{i,t}}\theta\,. \tag{11}$$

The marginal cost of stimulating customer activities at the expense of profit generation is equal to the marginal benefit (marginal q) of data, $V_{\delta_{i,t}}$, multiplied by data generation efficiency, θ.

In our model, data functions as productive capital, analogous to the role of capital in investment theories (Hayashi, 1982; Abel and Eberly, 1994), with a firm's product-design choices mirroring investment decisions, $x_{i,t}$. Specifically, the marginal q of a firm's data drives the choice of $x_{i,t}$, i.e., whether to prioritize customer engagement and data collection over monetization in

product design. A key distinction from the traditional investment models is that in our data economy, firms' investment decisions are interconnected through data sharing. Other firms' data stock, $\delta_{j,t}$, enters firm i's optimality condition (11) via the composite signal, $D_{i,t}$.

To sharpen the intuition, we specify the functional form of $\omega^m(e_{i,t})$ given by $\omega^m(e_{i,t}) = \ln(e_{i,t})$ and $C(D_{i,t}, x_{i,t})$ given by $C(D_{i,t}, x_{i,t}) = 1/(\phi_0 D_{i,t} - \phi_1 x_{i,t})$ that satisfy the properties in Section 4.1 and generate the profit function, $F(D_{i,t}, x_{i,t})$, through the optimal choice of marketing effort, $e_{i,t}$, given by (8).[28] Under these functional forms, we obtain closed-form solutions of the value function, $V^i\left(\delta_{i,t}, \{\delta_{j,t}\}_{j\neq i}\right)$, and product design (data investment) decision, $x_{i,t}$. We define

$$\hat{\rho} = \rho - \mu_\delta, \tag{12}$$

where, as defined in (4), μ_δ is negative, representing the depreciation of data (i.e., stale information on customer behavior is less informative). Therefore, $\hat{\rho}$ is essentially the user's cost of capital in traditional investment theory, i.e., the sum of discount rate (or required rate of return on capital) and depreciation rate. The following proposition summarizes the solution of value function.

Proposition 2 (Interconnected Firm Valuation) *Let* $\mathbf{D}_t = [D_{1,t}, ..., D_{i,t}, ... D_{N,t}]^\top$ *denote the column vector of all firms' signals on their customer. Firm* i*'s value at time* t *is given by*

$$V_{i,t} = \eta\delta_{i,t} + \frac{\eta\beta}{\hat{\rho}}\left\{\sum_{k=0}^{\infty}\left(\frac{\beta}{\hat{\rho}}\Gamma\right)^k \mathbf{D}_t\right\}_i + v_{i,0} = \eta\delta_{i,t} + \frac{\eta\beta}{\hat{\rho}}\left\{\left(\mathbf{I} - \frac{\beta}{\hat{\rho}}\Gamma\right)^{-1}\mathbf{D}_t\right\}_i + v_{i,0}. \tag{13}$$

where $v_{i,0}$ *and* η *are constant and defined in the appendix and the operator,* $\{\cdot\}_i$*, picks the* i*-th*

[28] Note that we do not introduce an additional parameter to scale $e_{i,t}$ in the logarithm function because the unit of marketing effort, $e_{i,t}$, can be freely interpreted.

element of a vector, under the following parameter condition,

$$\beta = \theta \left(\alpha + \frac{\phi_0}{\phi_1} \right) < \hat{\rho}, \tag{14}$$

that guarantees the convergence of $\sum_{k=1}^{\infty} \beta^k \Gamma^k$ *as the largest eigenvalue of* Γ *is one.*

Our solutions of firm valuation has several intuitive properties. The composite parameter β captures the overall growth rate of the data economy. As shown in (4), (5), and (6), $\theta\alpha$ reflects the autonomous growth of the data, as α translates the composite signal for firm i, $D_{i,t}$, into customer engagement $y_{i,t}$, and θ in turn translates $y_{i,t}$ into the growth of firm i's raw data, $\delta_{i,t}$. Additionally, firms can accumulate more data by adjusting product design. The choice of $x_{i,t}$ depends on the trade-off between current monetization and data accumulation as shown in (11). When ϕ_0 is high, firm i's signal on customers, $D_{i,t}$, significantly lowers the cost of acquiring paying customers, and when ϕ_1 is low, its choice of prioritizing customer engagement over monetization does not significant raise the cost of acquiring paying customers. Therefore, a high ratio of ϕ_0/ϕ_1 alleviates the tension between current profitability and data accumulation via customer engagement, allowing firm i to grow faster without sacrificing cash-flow generation (i.e., contributing positively to β).

We assume $\beta < \hat{\rho} = \rho - \mu_\delta$, so the discount rate, ρ, is greater than the net growth rate of data, $\beta + \mu_\delta$, where $\mu_\delta < 0$ captures data depreciation. This parameter condition is for the convergence of present value of cash flows, which is a standard in asset pricing. When calibrating the model, we interpret the discount rate to be a combination of interest rate and the intensity rate of an exogenous Poisson-arriving exit, and accordingly, measure ρ as the sum of these two components.[29]

The condition $\beta < \hat{\rho}$ also plays the role of ensuring the convergence of the network spillover effects, captured by $\sum_{k=0}^{\infty} \left(\frac{\beta}{\rho} \Gamma \right)^k$; otherwise, the infinite rounds of spillover effects explode.

[29]Under this interpretation, when a firm exits, we assume a new firm enters and inherits the exiting firm's $\delta_{i,t}$.

Mathematically, the geometric sequence of network propagation matrices converges when the network attenuation factor, $\beta/\hat{\rho}$, is greater than the largest eigenvalue of the network adjacency matrix, Γ, which is equal to one as Γ is right-stochastic (i.e., the row sums are equal to one, $\sum_{j=1}^{N} \gamma_{ij} = 1$).

As shown in (13), firm i's value depends positively on all firms' signals on their customers, $\mathbf{D}_t$. Such interconnectedness is captured by the data-sharing matrix, Γ. The infinite sum on the right side accounts for all the direct and indirect network propagation of firms' signals. The first term, $\Gamma\mathbf{D}_t$, reflects the direct (first degree) network externality and $\beta/\hat{\rho}$ is the network attenuation factor. The second-degree externality emerges as peer firms' signals in turn depend on signals from their connected firms. The network propagation mechanism depends crucially on the attenuation factor ($\beta/\hat{\rho} < 1$), with distant connections becoming increasingly weak, "discounted" by $(\beta/\hat{\rho})^k$.

The strength of network propagation (β) increases when a firm's signal on its customers is more effective in generating customer activities and cash flows (i.e., α is higher), when customer activities generate more data (i.e., θ is higher), and when data is very efficient in lowering marketing cost (i.e., ϕ_0 is higher). Overall, the more efficient the data economy is in generating data and turning data into profits, the more interconnected firms are in their valuation. As β increases and indirect connections degrees strengthen, firms' valuations become more correlated. In our empirical setting, ATT reduces θ (data collection efficiency), which weakens the interconnectedness.

Motivated by the empirical findings in Section 2, the model setup directly captures the comovement in firms' operational performances ($y_{i,t}$) and cash flows due to data sharing. After solving firms' valuation, we solve the stock return, $dR_{i,t} = dV_{i,t}/V_{i,t}$. The next corollary shows that data sharing induces return comovement and the ATT shock, i.e., a reduction of θ, reduces return comovement in line with our empirical findings in Section 2.

Corollary 1 (Stock Return Comovement) *The stock return of firm i at time t, $dR_{i,t} = dV_{i,t}/V_{i,t}$, has a correlation, $\rho_{r,i,j} = corr(dR_{i,t}, dR_{j,t})$, with firm j's stock return that is increasing in the*

dependence of firm i on firm j's data, $\gamma_{i,j}$, i.e., $\partial\rho_{r,i,j}/\partial\gamma_{i,j} > 0$. Additionally, we have $\frac{\partial^2\rho_{r,i,j}}{\partial\theta\partial\gamma_{i,j}} > 0$.

The decision makers in our model are the firm managers who rely on $D_{i,t}$, a sufficient statistic of its own data and other firms' data, to make product-design and marketing decisions. In reality, market participants may not be as informed about $D_{i,t}$ and the network structure of data sharing, so our valuation metric, V_t^i, may not map perfectly to stock-market valuation.

4.3 Calibration and model-implied comovements

Next, we calibrate our model to match the empirical patterns of comovements in firm performances induced by data sharing. The data-sharing connection between firm i and j, γ_{ij} (the ij-th element of Γ), is computed in Section 2, and we normalize Γ so that each row sums up to one. Therefore, the model takes the entire data-sharing network, shown in Figure 4, as an input. One unit of time is set to one quarter. For each firm on the network, we compute the volatility of quarterly growth rate of DAU (daily active users), our proxy for $\delta_{i,t}$, to pin down $\sigma_{\delta,i}$, capping the quarterly volatility at 40%. Firms with volatility less than 40% account for more than 98% of DAU in our sample. Next, we pin down κ and ζ by regressing firms' revenues on DAU to obtain $\zeta\kappa = 18$.[30] As a reminder, κ transforms cumulative customer engagement to the number of paying customers, and ζ represents profits per customer. Only the product, $\zeta\kappa$, enters into our model solutions.

We pin down a subset of parameters with external sources. Discount rate ρ has two components, interest rate and a Poisson intensity of firm exit. The former is set to 1%, and the latter set to 2% per quarter, consistent with the 8% annual exit rate in Jones and Kim (2018). Data depreciation rate is set to 7.5% per quarter following Veldkamp and Chung (2024), i.e., $\mu = -0.075$.

The rest of parameters we calibrate to replicate empirical findings in Section 2. Note that the

[30] Specifically, we regress total sales on DAU controlling for firm characteristics (total assets, cash, PP&E, and long-term debt) and industry-year fixed effects.

following parameters, α, θ, ϕ_0, and ϕ_1, form a composite parameter β. Model simulation requires β as a key input but does not require the values of these four parameters separately. We set β to 0.08. Moreover, we set ξ, the percentage reduction of θ caused by ATT, to 0.7, meaning that data collection from customer activities becomes 70% less efficient (see the law of motion of $\delta_{i,t}$ given by (4). Under $\beta = 0.08$ and $\xi = 0.7$, we simulate our model with the number of firms equal to that of our sample and firms interconnected via Γ as previously discussed. The simulation is done one hundred times, and each simulation is run for 20 quarters which is our sample length in Section 2.

With the simulated data, we run the regressions in Column (4) of Table 2, Column (4) of Table 3 and Column (4) of Table 4 and report the median estimates in the top panel of Table 6 alongside with the estimates from Table 2, 3 and 4. These regressions target the impact of data connectedness on comovement in firms' operational, financial, and stock-market performances respectively. Note that there are six regression coefficients but only two parameters, β and ξ, that we can choose to match these coefficients. The comparison in Table 6 shows that our model captures the comovements in firms' operational and stock-market performances reasonably well. In particular, the model successfully generates a strong positive association between firms' performance comovements and their data-sharing linkages, and the ATT shock weakens this mechanism.[31]

5 Product Design Dynamics

In Section 4, we develop a theoretical framework to capture the salient features of data economy that emerge from data sharing as documented in Section 2. Our model highlights that when designing products, firms face an intertemporal trade-off between monetization and data accumulation. The optimal product-design decision is characterized by a q-theory of data investment in Proposi-

[31] To estimate the DiD coefficients for the ATT shock, we introduce an unexpected change to θ in simulation. The ATT shock causes β to decline by 70%. As shown in (14) a 70% reduction of θ translates into a 70% decline of β.

tion 1. Next, we show that under data sharing, firms rationally mimic the product-design choices of one another. Such herding behavior in investment decisions is unique to data economy.

5.1 Intertemporal trade-off: monetization and data accumulation

The optimality condition for $x_{i,t}$ in Proposition 1 and the value function in Proposition 2 demonstrate that other firms' data, $\delta_{j,t}$, enters into firm i's choice of $x_{i,t}$ via the composite signal, $D_{i,t} = \sum_{j=1}^{N} \delta_{j,t}$. Next, we show that $x_{i,t}$ is increasing in $D_{i,t}$, and importantly, herding behavior in firms' product-design decisions emerges due to data sharing. ATT weakens the behavior.

Proposition 3 (Optimal Product Design) *Firm i prioritizes customer engagement and data accumulation over monetization (increases $x_{i,t}$) when $D_{i,t}$ is higher. The expected change of $x_{i,t}$, $\mathbb{E}_t[dx_{i,t}]$, is increasing in $x_{j,t}$, $j \neq i$, and the sensitivity is increasing in θ, i.e., $\frac{\partial \mathbb{E}_t[dx_{i,t}]}{\partial x_{j,t} \partial \theta} > 0$.*

In Lemma 1, we show that the cross-derivative of profit function is positive ($\frac{\partial^2 F_{i,t}}{\partial x_{i,t} \partial D_{i,t}} > 0$)—that is, other firms' data contribute to firm i's signal on its customers, $D_{i,t}$, and thus reduces firm i's marginal cost of stimulating user engagement and data collection at the expense of current profits. Intuitively, when firm i is more informed about its customers, its marketing efforts become more effective in generating profits in spite of a product design that downplays monetization. Therefore, other firms collecting data reduces firm i's marginal cost of accumulating data (and downplaying monetization). The strength of such spillover effect is given by γ_{ij}, i.e., $\delta_{j,t}$'s coefficient in $D_{i,t}$.

Consider an increase in other firms' current choice of $x_{j,t}$, which leads to an increase in their data stock, $\delta_{j,t}$, and an increase in firm i's signal on its customers, $D_{i,t}$. Then from t to $t + dt$, firm i's choice of $x_{i,t}$ increases, resulting in a "herding" or cross-firm momentum in prioritizing data accumulation over monetization in product design. Thus, a positive data investment externality exists, which is sharp contrast with the investment dynamics of traditional firms whose investment

is likely to crowd out other firms' investment (for example, by raising the prices of investment inputs and financing costs). In the next subsection, we provide empirical evidence.

The product-design dynamics also suggests that waves of active data collection or monetization emerge in the data economy. Shocks to one firm's data stock are propagated to other firms through data sharing, and as a result, other firms marginal cost of data investment are affected. Positive shocks to one firm leads to more data investments by other firms, so all firms in the economy tend to prioritize customer engagement and data collection over monetization. In contrast, negative shocks to one firm are propagated through the data-sharing network, causing other firms to respond by prioritizing monetization in their product design at the expense of customer engagement.

Overall, due to the positive spillover effect (one firm's data accumulation reduces other firms' cost of data investment), our model features under-investment in data accumulation, that is firms' choices of $x_{i,t}$ are below those deemed optimal by the planner that maximizes all firms' values.[32] Under-investment intensifies during a monetization wave that is often triggered by negative shocks (e.g., cyberattack) to one or several firms' data stock.

Proposition 3 shows that by reducing θ, ATT weakens the cross-firm momentum in product-design decisions, i.e., $\frac{\partial \mathbb{E}_t[dx_{i,t}]}{\partial x_{j,t} \partial \theta} > 0$. This is an unintended consequence of ATT that has not been studied before. Beyond the impact on herding in product design, reducing θ directly reduces firms' incentive to acquire data (i.e., decreases data marginal q) and encourages firms to prioritize monetization, in line with the findings in Kesler (2023). Intuitively, ATT dampens the self-perpetuating dynamics of data accumulation (see the data dynamics block of our model in Figure 7).[33]

Proposition 4 (ATT and Product Design) *Reducing θ reduces the data marginal q, $\partial V_{i,t}/\partial \delta_{i,t}$, and thus causes all firms to prioritize monetization over customer engagement (i.e., $x_{i,t}$ decreases).*

[32] Note that when computing the planner's solution, we do not consider other stake holders' welfare, including those of the customers and employees. The planner's objective is to maximize the sum of all firms' value.

[33] Specifically, the link from $y_{i,t}$, customer activities, to future $\delta_{i,t}$ is weakened.

5.2 Evidence on interconnected product design choices

We test the unique prediction of our model on product-design dynamics in the data economy—the herding behavior in firms' product-design decisions in Proposition 3. Moreover, we show that in line with Proposition 4, ATT has the unintended consequence of weakening such herding behavior.

To examine whether firms' product-design decisions (prioritizing monetization vs. customer engagement) and how they are influenced by their peers, we design the following empirical strategy. First, we classify payment SDKs as directly related to monetization, while SDKs facilitating data security and customer support are related to customer engagement. Examples of those SDKs are provided in Appendix A. For each of the following SDK categories–payment, security, and customer support–we calculate: 1) the number of unique SDKs used by a firm, denoted as $X_{i,t}$, 2) the change in the number of unique SDKs used by a firm, denoted as $\Delta X_{i,t}$, and 3) the weighted sum of the number of unique SDKs used by peer firms that share data with the focal firm, where the weight is the pairwise data connectedness, specified as follows

$$X_{-i,t-1} = \sum_{j,j\neq i} \rho_{i,j}^{\text{data}} X_{j,t-1}.$$

Importantly, the third measure $X_{-i,t-1}$ captures peer firms' strategies.

We then estimate the following equation separately for different SDK categories:

$$\Delta X_{i,t} = \alpha + \beta_1 X_{-i,t-1} + \beta_2 \text{ATT} \times X_{-i,t-1} + \beta_3 X_{i,t-1} + \mathbf{C_{it}}\beta_4 + \iota_{kt} + \varepsilon_{i,t}, \tag{15}$$

where $\mathbf{C_{it}}$ includes firm-level controls such as firm size (log of assets), long-term debt to assets, and tangible assets to total assets. We also include industry-year fixed effects (ι_{kt}). Standard errors are clustered at the firm level. The coefficient of β_1 corresponds to the model prediction on herding

behavior in firms' product-design decisions in Proposition 3 and the coefficient β_2 corresponds to the impact of ATT in Proposition 4. The estimation results are presented in Table 7, with Panel A focusing on monetization SDKs and Panel B on SDKs related to customer engagement.

Consistent with the model's prediction, the changes in firms' product design positively load on their peers' product-design choices from the previous period across different functionalities of the SDKs. Additionally, the coefficient on the interaction term, β_2 for ATT $\times X_{-i,t-1}$, is negative. This indicates that the herding behavior emerges from data sharing, and ATT weakens this channel.

Using the simulated data from Section 4.3, we run the same regressions, and in Table 8, we compare the model-implied regression coefficients with those estimated from data. The model-generated cross-firm herding in product-design choices is in line with that observed in data. Moreover, the model also generates a weakened herding behavior after the ATT shock in line with our empirical findings. Therefore, even though the product-design dynamics are not targeted when we calibrate the model, the model generates patterns that closely resemble those observed in the data.

6 Systemically Important Firms in Data Economy

Data sharing generates interdependence across firms. To comprehensively evaluate firms' systemic importance in the data economy, we use our dynamic network model to develop valuation-based metrics, which incorporate firm size heterogeneity, the topology of data-sharing network, and spillover effects from both direct and indirect connections and over multiple time horizons.

The reduced-form evidence in Section 3.1 and 5.2 is based on firms' direct connections through data sharing. This strategy omits two critical aspects of network externalities, the higher-order externalities and persistent impact across multiple time horizons. Impact of a firm's data on peer firms is likely to transmit further to their data-sharing counterparts, resulting in higher-degree

externalities. Additionally, as illustrated in Figure 7, data generates self-reinforcing growth, so a firm's data has a persistent effect on itself and its connected peers by affecting the trajectory of data growth. In the following, we demonstrate that our valuation metrics comprehensively summarize the network externalities of higher orders and over multiple time horizons.

Network-augmented Gordon growth formula. A firm's valuation is the present value of cash flows that reflect the productivity of the firm's own data and the dependence on other firms' data across different time horizons. Any variation in a peer firm j's data stock, $\delta_{j,t}$, through the composite signal, $D_{i,t} = \sum_{j=1}^{N} \gamma_{ij}\delta_{j,t}$, affects firm i's product-design decision, $x_{i,t}$, customer engagement (i.e., $y_{i,t}$ in (6)), and the growth of firm i's data, $\delta_{i,t}$, given by (4). The impact can be traced through the diagram in Figure 7. Importantly, the impact is persistent. Given the geometric growth path of data stock, any variation in its current value, $\delta_{i,t}$, shifts the whole trajectory. Moreover, as the variation in $\delta_{j,t}$ transmits through firm i to other firms, it generates a second-degree spillover.

Next, we show that in the absence of data sharing, the valuation formula is reduced to a standard Gorton growth formula. Under data sharing, the firm-level growth is replaced with "community growth" of the whole data economy into the valuation framework, and the topology of data sharing, given by the network adjacency matrix, Γ, plays a crucial role in determining the overall community growth and its contribution to firm i's valuation.

Proposition 5 (Network-Augmented Gordon Growth Formula) *Let $\bar{\delta}_t$ denote the column vector of all firms' data stock, $\bar{\delta}_t = [\delta_{1,t}, ..., \delta_{N,t}]^\top$. Firm i's value can be written as*

$$V_{i,t} = \zeta\kappa \left\{ (\hat{\rho}\mathbf{I} - \beta\Gamma)^{-1} \bar{\delta}_t \right\}_i + v_{i,0}. \tag{16}$$

where $v_{i,0}$ is constant and the operator, $\{\cdot\}_i$, picks the i-th element of a vector. In the absence of

data sharing, i.e., $\Gamma = \mathbf{I}$*, firm* i*'s value is given by*

$$V_{i,t} = \left(\frac{\zeta \kappa}{\hat{\rho} - \beta} \right) \delta_{i,t} + v_{i,0}. \tag{17}$$

Data sharing generates a network-augmented Gordon Growth formula. Given all firms' data stocks at time t, $\bar{\delta}_t = [\delta_{1,t}, ..., \delta_{N,t}]^\top$, the linear operator, $(\hat{\rho}\mathbf{I} - \beta\Gamma)^{-1}$, generates the expected growth paths for all firms and discounts the growth paths with $\hat{\rho}$. The rotated data vector $(\hat{\rho}\mathbf{I} - \beta\Gamma)^{-1} \bar{\delta}_t$ represents the persistent impact of data on firms' cash flows at multiple time horizons and the spillover of data from one firm to another. The i-the element belongs to firm i. When data sharing is eliminated (i.e., $\Gamma = \mathbf{I}$), $(\hat{\rho}\mathbf{I} - \beta\Gamma)^{-1}$ in (16) is replaced by $1/(\hat{\rho} - \beta)$ in (17). The coefficient of $\delta_{i,t}$ in (17), $\zeta\kappa/(\hat{\rho} - \beta)$, is the present value of cash flows generated per unit of $\delta_{i,t}$ under $\Gamma = \mathbf{I}$. Therefore, in the absence of data sharing, firm i's valuation is given by the standard Gordon growth formula (17) with $\delta_{i,t}$, the data stock, as a scaling factor.

Our valuation framework is forward-looking and properly accounts for network externalities. This stands in contrast to the traditional cost-based method of valuing intangible capital. Broadly speaking, intangible capital exhibits positive network externalities through knowledge spillover across firms and the scope of usage across firms and industries. Data is one salient example. We provide a direct measure of spillover effects, i.e., the data-sharing network Γ, and incorporate it into a valuation framework to evaluate its economic significance.

Valuation centrality. Next, we develop a method to identify systemically important firms in the data economy. Aggregating the valuation of all firms, we obtain

$$\bar{V}_t = \zeta\kappa \mathbf{1}^\top (\hat{\rho}\mathbf{I} - \beta\Gamma)^{-1} \bar{\delta}_t + \sum_i v_{i,0}, \tag{18}$$

which connects the column vector of all firms' data stock, $\bar{\delta}_t$, to the value of aggregate cash flows across all time horizons. Any variation in a particular firm's data stock is transmitted to cash flows across all firms at all horizons through the network-augmented growth matrix,

$$\mathbf{1}^\top \left(\hat{\rho}\mathbf{I} - \beta\Gamma\right)^{-1} = \frac{1}{\hat{\rho}}\mathbf{1}^\top \sum_{k=0}^{\infty} \left(\frac{\beta}{\hat{\rho}}\Gamma\right)^k. \tag{19}$$

Each firm's contribution to the aggregate value of cash flows is given by

$$\zeta\kappa\{\mathbf{1}^\top \left(\hat{\rho}\mathbf{I} - \beta\Gamma\right)^{-1}\}_{.i}\delta_{i,t} + v_{i,0}, \tag{20}$$

where $\{\cdot\}_{.i}$ picks the i-th column of a matrix. $\{\mathbf{1}^\top \left(\hat{\rho}\mathbf{I} - \beta\Gamma\right)^{-1}\}_{.i}$ encodes all the routes of data spillover from firm i to other firms and summarizes such impact across all time horizons. Our metric of valuation contribution traces the flow of firm i's data through the whole economy, thus providing a comprehensive account of data deployment across different firms. Our focus on the scope of data usage across firms echoes Crouzet et al. (2024) who emphasize the scope of intangible capital usage across different divisions within firms and by competing imitators.

Following Ballester et al. (2006) and Denbee et al. (2021), we define *valuation key player* as the firm who makes the largest contribution to the aggregate value of cash flows, i.e.,

$$\text{VKP} = \arg\max_i \ \zeta\kappa\{\mathbf{1}^\top \left(\hat{\rho}\mathbf{I} - \beta\Gamma\right)^{-1}\}_{.i}\delta_{i,t} + v_{i,0}. \tag{21}$$

In Panel B of Figure 8, we report the valuation contribution from the top 50 firms ranked by their DAU and compare it against firms' DAU in Panel A. In both panels, we rank firms by DAU so the bar chart in Panel A exhibits a monotonically decreasing pattern, and we normalize firms' values by that of the first firm. Comparing the two panels reveals that firms' valuation contribution,

which takes into account the data spillover effects of all degrees of network propagation and across all time horizons, can differ significantly from firm size. Therefore, to understand the economic implications of data sharing, it is critical to trace the network of data flows.

A firm's contribution to the value of aggregate cash flows given by (20) can also be decomposed into rounds of network propagation:

$$\frac{\zeta\kappa}{\hat{\rho}}\left\{\mathbf{1}^{\top}\sum_{k=0}^{\infty}\left(\frac{\beta}{\hat{\rho}}\right)^{k}\Gamma^{k}\right\}_{.i}\delta_{i,t}+v_{i,0}. \tag{22}$$

At time t, the stock of firm i's raw data is $\delta_{i,t}$. By contributing to its own and other firms' signals, variation of $\delta_{i,t}$ permeates across the data-sharing network generating direct ($k=1$) and indirect ($k>1$) spillover effects and, through the self-reinforcing data dynamics as shown in Figure 7, such impact persists into the future. In Figure 9, we truncate k at different values, denoted by K, starting from $K=0$, and report firms' valuation contribution. At $K=0$, firm i's contribution to the value of aggregate cash flows is given by $\frac{\zeta\kappa}{\hat{\rho}}\delta_{i,t}$, which shuts down the propagation of data to other firms and over time. In Panel B and C, we consider $K=3$ and $K=\infty$, respectively. As K increases, valuation contribution converges to the equilibrium value.

Note that a firm's contribution to the value of aggregate cash flows differs from its own valuation. In Figure 10, we report the ratio of valuation contribution to firm's own valuation:

$$\frac{\zeta\kappa\{\mathbf{1}^{\top}\left(\hat{\rho}\mathbf{I}-\beta\Gamma\right)^{-1}\}_{.i}\delta_{i,t}+v_{i,0}}{\zeta\kappa\mathbf{e}_{i}^{\top}\left(\hat{\rho}\mathbf{I}-\beta\Gamma\right)^{-1}\bar{\delta}_{t}+v_{i,0}}. \tag{23}$$

In Panel A, we solve the model and compute the ratios under the pre-ATT value of β, and in Panel B, we compute the post-ATT ratios. In both panels, firms are ranked by DAU. Pre-ATT, contribution from Meta (formerly Facebook) to the value of aggregate cash flows in the data economy is more than 2.5x Meta's own valuation, and Alphabet (formerly Google) has an even higher ra-

tio above 5.0x. After the introduction of ATT that curtails the cross-firm data flows, Meta's and Alphabet's ratios declined to 1.2x and 2.4x respectively.

7 Conclusion

Data's non-rival nature and externalities make it a uniquely powerful asset. One firm's data can be used simultaneously by other firms. This paper uncovers a network of inter-firm data flows, facilitated by data analytics software. Firms collect data on customers and share with one another for customer profiling, which is critical for enhancing firms' operational efficiency, improving customer engagement, and supporting further data collection in a self-reinforcing cycle.

We document that data sharing drives strong comovements in operational, financial, and stock-market performances among data-connected firms. Motivated by these empirical findings, we develop a dynamic network model of data economy that captures the interconnected dynamics of data collection, sharing, and utilization, providing insights into the economic implications of policy interventions (e.g., ATT). In addition, our model reveals a novel feature of data-driven firms that is supported by evidence: their product-design decisions exhibit herding. Finally, the model allows us to identify systemically important firms whose critical positions in the data-sharing network disproportionately influence the data economy. These findings highlight the need to consider the network structure of data sharing when evaluating the role of data as a productive asset.

References

Abel, A. B. and J. C. Eberly (1994). A unified model of investment under uncertainty. *The American Economic Review 84*(5), 1369–1384.

Abis, S. and L. Veldkamp (2023). The Changing Economics of Knowledge Production. *The Review of Financial Studies 37*(1), 89–118.

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2022). Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics 14*(4), 218–56.

Acharya, V. V., L. H. Pedersen, T. Philippon, and M. Richardson (2016). Measuring Systemic Risk. *The Review of Financial Studies 30*(1), 2–47.

Adrian, T. and M. K. Brunnermeier (2016). Covar. *American Economic Review 106*(7), 1705–41.

Akey, P., S. Lewellen, I. Liskovich, and C. Schiller (2023). Hacking corporate reputations. *Rotman School of Management Working Paper* (3143740).

Alcobendas, M., S. Kobayashi, K. Shi, and M. Shum (2023). The impact of privacy protection on online advertising markets. *Available at SSRN 3782889*.

Ali, U. and D. Hirshleifer (2020). Shared analyst coverage: Unifying momentum spillover effects. *Journal of Financial Economics 136*(3), 649–675.

Argente, D., S. Moreira, O. Ezra, and V. Venky (2021). Scalable expertise. Working paper.

Aridor, G., Y.-K. Che, B. Hollenbeck, M. Kaiser, and D. McCarthy (2024). Evaluating the impact of privacy regulation on e-commerce firms: Evidence from apple's app tracking transparency. *Available at SSRN*.

Aridor, G., Y.-K. Che, and T. Salz (2020). The economic consequences of data privacy regulation: Empirical evidence from GDPR. Technical report, National Bureau of Economic Research.

Asriyan, V., L. Laeven, A. Martin, A. Van der Ghote, and V. Vanasco (2024). Falling interest rates and credit reallocation: Lessons from general equilibrium. *The Review of Economic Studies*, Forthcoming.

Auer, R. A., A. A. Levchenko, and P. Sauré (2019). International inflation spillovers through input linkages. *Review of Economics and Statistics 101*(3), 507–521.

Bai, J., A. Krishnamurthy, and C.-H. Weymuller (2018). Measuring liquidity mismatch in the banking sector. *The Journal of Finance 73*(1), 51–93.

Ballester, C., A. Calvó-Armengol, and Y. Zenou (2006). Who's who in networks. Wanted: The key player. *Econometrica 74*, 1403–1417.

Ballester, C., A. Calvo-Armengol, and Y. Zenou (2006). Who's who in networks. Wanted: The key player. *Econometrica 74*, 1403–1417.

Barrot, J.-N. and J. Sauvagnat (2016). Input specificity and the propagation of idiosyncratic shocks in production networks. *The Quarterly Journal of Economics 131*(3), 1543–1592.

Baxter, M. and M. A. Kouparitsas (2005). Determinants of business cycle comovement: a robust analysis. *Journal of Monetary Economics 52*(1), 113–157.

Begenau, J., M. Farboodi, and L. Veldkamp (2018). Big data in finance and the growth of large firms. *Journal of Monetary Economics 97*, 71–87.

Benoit, S., J.-E. Colliard, C. Hurlin, and C. Pérignon (2016). Where the Risks Lie: A Survey on Systemic Risk. *Review of Finance 21*(1), 109–152.

Bergemann, D. and A. Bonatti (2019). The Economics of Social Data: An Introduction. Cowles Foundation Discussion Papers 2171R, Cowles Foundation for Research in Economics, Yale University.

Bessen, J. E., S. M. Impink, L. Reichensperger, and R. Seamans (2020). GDPR and the importance of data to AI startups. Working paper, New York University, Boston University.

Bhandari, A. and E. R. McGrattan (2021). Sweat Equity in U.S. Private Business. *The Quarterly Journal of Economics 136*(2), 727–781.

Bian, B., X. Ma, and H. Tang (2021). The supply and demand for data privacy: Evidence from mobile apps. *Available at SSRN*.

Bian, B., M. Pagel, H. Tang, and D. Raval (2023). Consumer surveillance and financial fraud. Technical report, National Bureau of Economic Research.

Billio, M., M. Getmansky, A. W. Lo, and L. Pelizzon (2012). Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics 104*(3), 535–559.

Binns, R., U. Lyngs, M. Van Kleek, J. Zhao, T. Libert, and N. Shadbolt (2018). Third party tracking in the mobile ecosystem. In *Proceedings of the 10th ACM Conference on Web Science*, pp. 23–31.

Bloom, N., M. Schankerman, and J. Van Reenen (2013). Identifying technology spillovers and product market rivalry. *Econometrica 81*(4), 1347–1393.

Blume, L. E., W. A. Brock, S. N. Durlauf, and R. Jayaraman (2015). Linear social interactions models. *Journal of Political Economy 123*(2), 444–496.

Boehm, C. E., A. Flaaen, and N. Pandalai-Nayar (2019). Input linkages and the transmission of shocks: Firm-level evidence from the 2011 tōhoku earthquake. *Review of Economics and Statistics 101*(1), 60–75.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009). Identification of peer effects through social networks. *Journal of Econometrics 150*(1), 41–55.

Calvó-Armengol, A., E. Patacchini, and Y. Zenou (2009). Peer effects and social networks in education. *The Review of Economic Studies 76*(4), 1239–1267.

Carvalho, V. M., M. Nirei, Y. U. Saito, and A. Tahbaz-Salehi (2021). Supply chain disruptions: Evidence from the great east japan earthquake. *The Quarterly Journal of Economics 136*(2), 1255–1321.

Choi, J. P., D.-S. Jeon, and B.-C. Kim (2019). Privacy and personal data collection with information externalities. *Journal of Public Economics 173*, 113–124.

Cohen, L. and A. Frazzini (2008). Economic links and predictable returns. *The Journal of Finance 63*(4), 1977–2011.

Cohen-Cole, E., A. Kirilenko, and E. Patacchini (2014). Trading networks and liquidity provision. *Journal of Financial Economics 113*(2), 235–251.

Cong, L. W., D. Xie, and L. Zhang (2021). Knowledge accumulation, privacy, and growth in a data economy. *Management Science 67*(10), 6480–6492.

Corhay, A., K. Hu, J. Li, J. Tong, and C.-Y. Tsou (2024). Data, markups, and asset prices. Working paper.

Crosignani, M., M. Macchiavelli, and A. F. Silva (2023). Pirates without borders: The propagation of cyberattacks through firms' supply chains. *Journal of Financial Economics 147*(2), 432–448.

Crouzet, N., J. Eberly, A. Eisfeldt, and D. Papanikolaou (2024). Intangible capital, firm scope, and growth. Working paper.

Crouzet, N., J. C. Eberly, A. L. Eisfeldt, and D. Papanikolaou (2022). The economics of intangible capital. *Journal of Economic Perspectives 36*(3), 29–52.

Daniel, K., M. Grinblatt, S. Titman, and R. Wermers (1997). Measuring mutual fund performance with characteristic-based benchmarks. *The Journal of Finance 52*(3), 1035–1058.

de Paula, A. (2017). Econometrics of network models. In *Advances in Economics and Econometrics: Theory and Applications, Eleventh World Congress*, pp. 268–323. Cambridge University Press.

Demir, B., B. Javorcik, T. K. Michalski, and E. Ors (2024). Financial constraints and propagation of shocks in production networks. *Review of Economics and Statistics 106*(2), 437–454.

Demirer, M., D. J. Jiménez Hernández, D. Li, and S. Peng (2024). Data, privacy laws and firm production: Evidence from the GDPR. Working Paper 32146, National Bureau of Economic Research.

Denbee, E., C. Julliard, Y. Li, and K. Yuan (2021). Network risk and key players: A structural analysis of interbank liquidity. *Journal of Financial Economics 141*(3), 831–859.

Di Giovanni, J. and A. A. Levchenko (2010). Putting the parts together: trade, vertical linkages, and business cycle comovement. *American Economic Journal: Macroeconomics 2*(2), 95–124.

Diebold, F. X. and K. Yılmaz (2014). On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics 182*(1), 119–134.

Dou, W. W., Y. Ji, D. Reibstein, and W. Wu (2021). Inalienable customer capital, corporate liquidity, and stock returns. *The Journal of Finance 76*(1), 211–265.

Duarte, F. and T. M. Eisenbach (2021). Fire-sale spillovers and systemic risk. *The Journal of Finance 76*(3), 1251–1294.

Eeckhout, J. and L. Veldkamp (2022). Data and markups: A macro-finance perspective. Working Paper 30022, National Bureau of Economic Research.

Eisfeldt, A. L., B. Herskovic, and S. Liu (2023). Interdealer price dispersion. Working paper, UCLA.

Eisfeldt, A. L., B. Herskovic, S. Rajan, and E. Siriwardane (2022). OTC Intermediaries. *The Review of Financial Studies 36*(2), 615–677.

Eisfeldt, A. L. and D. Papanikolaou (2013). Organization capital and the cross-section of expected returns. *The Journal of Finance 68*(4), 1365–1406.

Ewens, M., R. H. Peters, and S. Wang (2024). Measuring intangible capital with market prices. *Management Science 2024*(0).

Fajgelbaum, P. D., E. Schaal, and M. Taschereau-Dumouchel (2017). Uncertainty traps*. *The Quarterly Journal of Economics 132*(4), 1641–1692.

Farboodi, M., R. Mihet, T. Philippon, and L. Veldkamp (2019). Big data and firm dynamics. *AEA Papers and Proceedings 109*, 38–42.

Farboodi, M., D. Singal, L. Veldkamp, and V. Venkateswaran (2024). Valuing financial data. *The Review of Financial Studies*, hhae034.

Farboodi, M. and L. Veldkamp (2020). Long-run growth of financial data technology. *American Economic Review 110*(8), 2485–2523.

Farboodi, M. and L. Veldkamp (2021). A model of the data economy. Working Paper 28427, National Bureau of Economic Research.

Florackis, C., C. Louca, R. Michaely, and M. Weber (2022). Cybersecurity risk. *The Review of Financial Studies 36*(1), 351–407.

Fogli, A. and L. Veldkamp (2021). Germs, social networks, and growth. *The Review of Economic Studies 88*(3), 1074–1100.

Frésard, L., G. Hoberg, and G. M. Phillips (2020). Innovation activities and integration through vertical acquisitions. *The Review of Financial Studies 33*(7), 2937–2976.

Glaeser, E. L. and J. Scheinkman (2000). Non-market interactions. Working Paper 8053, National Bureau of Economic Research.

Goldberg, S., G. Johnson, and S. Shriver (2019). Regulating privacy online: The early impact of the GDPR on european web traffic & e-commerce outcomes. *Available at SSRN 3421731*.

Gourio, F. and L. Rudanko (2014). Customer capital. *The Review of Economic Studies 81*(3), 1102–1136.

Graham, B. S. (2008). Identifying social interactions through conditional variance restrictions. *Econometrica 76*(3), 643–660.

Greenwood, R., A. Landier, and D. Thesmar (2015). Vulnerable banks. *Journal of Financial Economics 115*, 471–485.

Hayashi, F. (1982). Tobin's marginal q and average q: A neoclassical interpretation. *Econometrica 50*(1), 213–224.

Herskovic, B. (2018). Networks in production: Asset pricing implications. *The Journal of Finance 73*(4), 1785–1818.

Herskovic, B., B. Kelly, H. Lustig, and S. Van Nieuwerburgh (2020). Firm volatility in granular networks. *Journal of Political Economy 128*(11), 4097–4162.

Hoberg, G. and G. Phillips (2010). Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies 23*(10), 3773–3811.

Hoberg, G. and G. Phillips (2016). Text-based network industries and endogenous product differentiation. *Journal of political economy 124*(5), 1423–1465.

Hoberg, G. and G. M. Phillips (2018). Text-based industry momentum. *Journal of Financial and Quantitative Analysis 53*(6), 2355–2388.

Hsieh, C.-T. and E. Rossi-Hansberg (2023). The industrial revolution in services. *Journal of Political Economy Macroeconomics 1*(1), 3–42.

Ichihashi, S. (2020). Online privacy and information disclosure by consumers. *American Economic Review 110*(2), 569–95.

Ichihashi, S. (2021). The economics of data externalities. *Journal of Economic Theory 196*, 105316.

Imbs, J. (2004). Trade, finance, specialization, and synchronization. *Review of economics and statistics 86*(3), 723–734.

Jaffe, A. (1986). Technological opportunity and spillovers of r&d: Evidence from firms' patents, profits, and market value. *American Economic Review 76*(5), 984–1001.

Janssen, R., R. Kesler, M. E. Kummer, and J. Waldfogel (2022). GDPR and the lost generation of innovative apps. Technical report, National Bureau of Economic Research.

Jia, J., G. Z. Jin, and L. Wagman (2021). The short-run effects of the general data protection regulation on technology venture investment. *Marketing Science 40*(4), 661–684.

Jiang, Z. and R. Richmond (2021). Origins of international factor structures. Working paper, NYU Stern and Northwestern Kellogg.

Jin, G. Z., Z. Liu, and L. Wagman (2024). The gdpr and sdk usage in android mobile apps. *Law & Economics Center at George Mason University Scalia Law School Research Paper Series Forthcoming*.

Johnson, G. (2022). Economic research on privacy regulation: Lessons from the GDPR and beyond.

Johnson, G. A., S. K. Shriver, and S. G. Goldberg (2023). Privacy and market concentration: Intended and unintended consequences of the GDPR. *Management Science 69*(10), 5695–5721.

Jones, C. I. and J. Kim (2018). A schumpeterian model of top income inequality. *Journal of Political Economy 126*(5), 1785–1826.

Jones, C. I. and C. Tonetti (2020). Nonrivalry and the economics of data. *American Economic Review 110*(9), 2819–58.

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2021). Measuring technological innovation over the long run. *American Economic Review: Insights 3*(3), 303–20.

Kesler, R. (2023). The impact of apple's app tracking transparency on app monetization. *Available at SSRN 4090786*.

Kogan, L., D. Papanikolaou, A. Seru, and N. Stoffman (2017). Technological Innovation, Resource Allocation, and Growth. *The Quarterly Journal of Economics 132*(2), 665–712.

Kraft, L., B. Skiera, and T. Koschella (2023). Economic impact of opt-in versus opt-out requirements for personal data usage: the case of apple's app tracking transparency (att). *Available at SSRN 4598472*.

Li, D. and H.-T. T. Tsai (2022). Mobile apps and targeted advertising: Competitive effects of data sharing. *Available at SSRN 4088166*.

Li, Y., Y. Li, and H. Sun (2023). The network structure of money multiplier. Working paper, Columbia University, Federal Reserve Board, and University of Washington.

Liu, E. and S. Ma (2021). Innovation networks and r&d allocation. Technical report, National Bureau of Economic Research.

Liu, E., S. Ma, and L. Veldkamp (2023). Data sales and data dilution. Working paper, Columbia University, Princeton University, Yale University.

McGrattan, E. R. and E. C. Prescott (2009). Openness, technology capital, and development. *Journal of Economic Theory 144*(6), 2454–2476. Dynamic General Equilibrium.

McGrattan, E. R. and E. C. Prescott (2010). Technology capital and the us current account. *American Economic Review 100*(4), 1493–1522.

Menzly, L. and O. Ozbas (2010). Market segmentation and cross-predictability of returns. *The Journal of Finance 65*(4), 1555–1580.

Ordoñez, G. L. (2013). Fragility of reputation and clustering of risk-taking. *Theoretical Economics 8*(3), 653–700.

Ozdagli, A. and M. Weber (2017). Monetary policy through production networks: Evidence from the stock market. Working Paper 23424, National Bureau of Economic Research.

Parsons, C. A., R. Sabbatucci, and S. Titman (2020). Geographic lead-lag effects. *The Review of Financial Studies 33*(10), 4721–4770.

Peters, R. H. and L. A. Taylor (2017). Intangible capital and the investment-q relation. *Journal of Financial Economics 123*(2), 251–272.

Peukert, C., S. Bechtold, M. Batikas, and T. Kretschmer (2022). Regulatory spillovers and data governance: Evidence from the GDPR. *Marketing Science 41*(4), 746–768.

Varian, H. (2018). Artificial intelligence, economics, and industrial organization. Working Paper 24839, National Bureau of Economic Research.

Veldkamp, L. (2023). Valuing Data as an Asset. *Review of Finance 27*(5), 1545–1562.

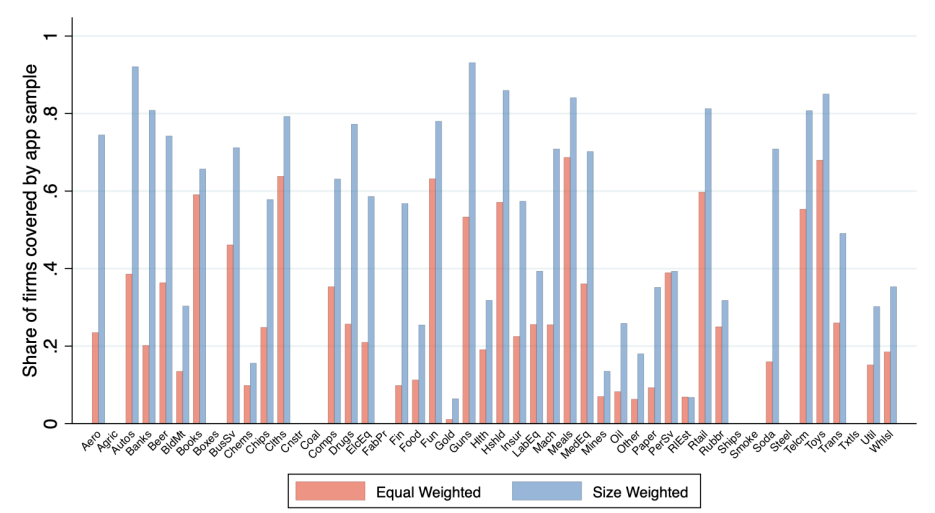
Veldkamp, L. and C. Chung (2024). Data and the aggregate economy. *Journal of Economic Literature 62*(2), 458–484.

Veldkamp, L. L. (2005). Slow boom, sudden crash. *Journal of Economic Theory 124*(2), 230–257.

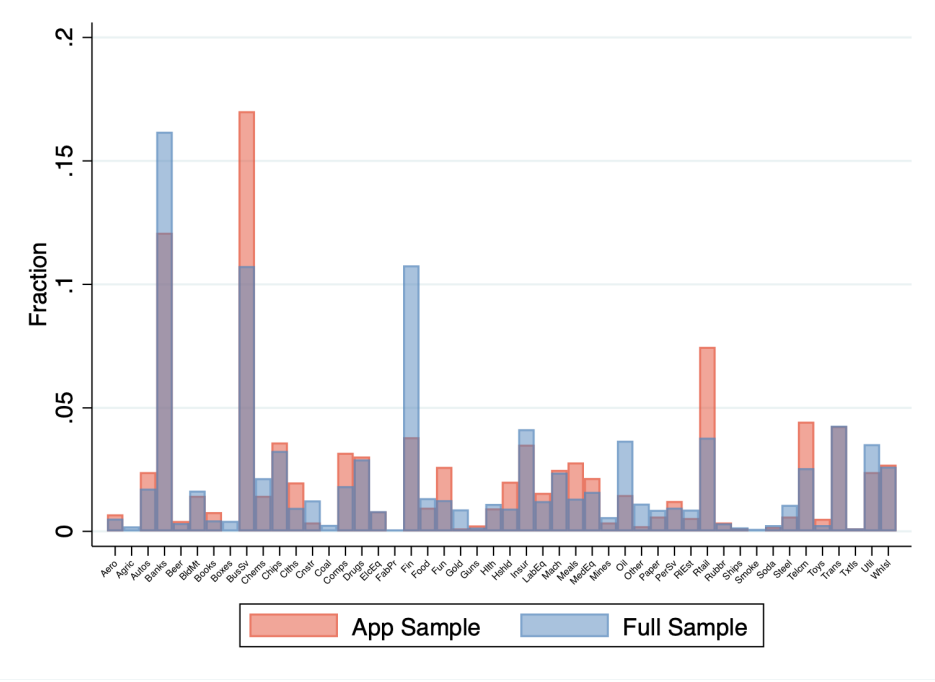
Wernerfelt, N., A. Tuchman, B. T. Shapiro, and R. Moakler (2024). Estimating the value of offsite tracking data to advertisers: Evidence from meta. *Marketing Science*.

FIGURE 1: **Firms in the Data Network vs. Compustat Universe – Industry**

Panel A. Share of data-reliant firms within industry



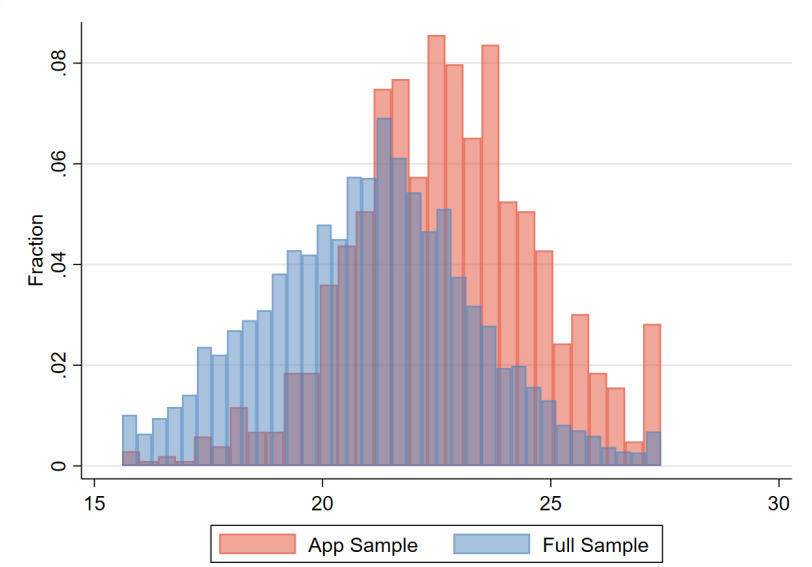
Panel B. Share of data-reliant/Compustat firms across industries



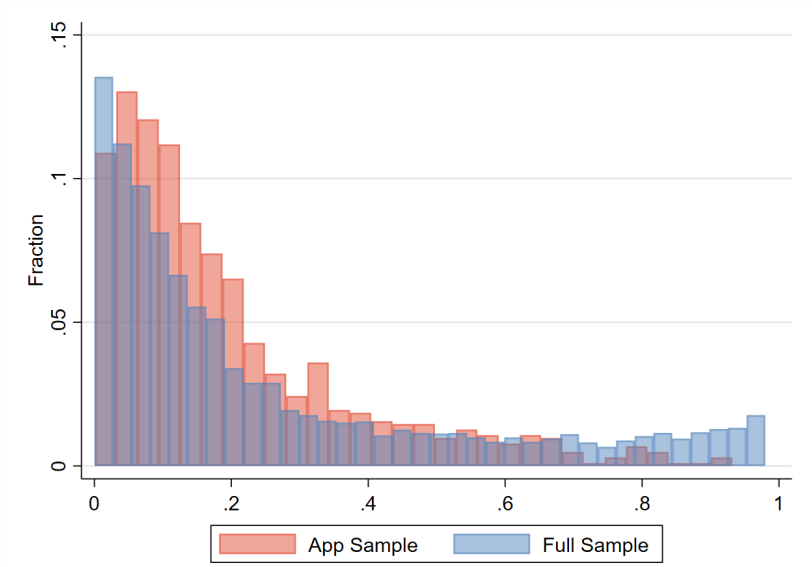
NOTE.—Figure 1 compares the industry distribtion of firms in the data network ("App Sample") with those in the broader Compustat universe ("Full Sample"). Panel A shows the proportion of data-reliant firms within each industry Fama-French 48 industry classification and Panel B shows the distribution of firms across industries.

FIGURE 2: **Firms in the Data Network vs. Compustat Universe – Size and Sales**

Panel A. Size

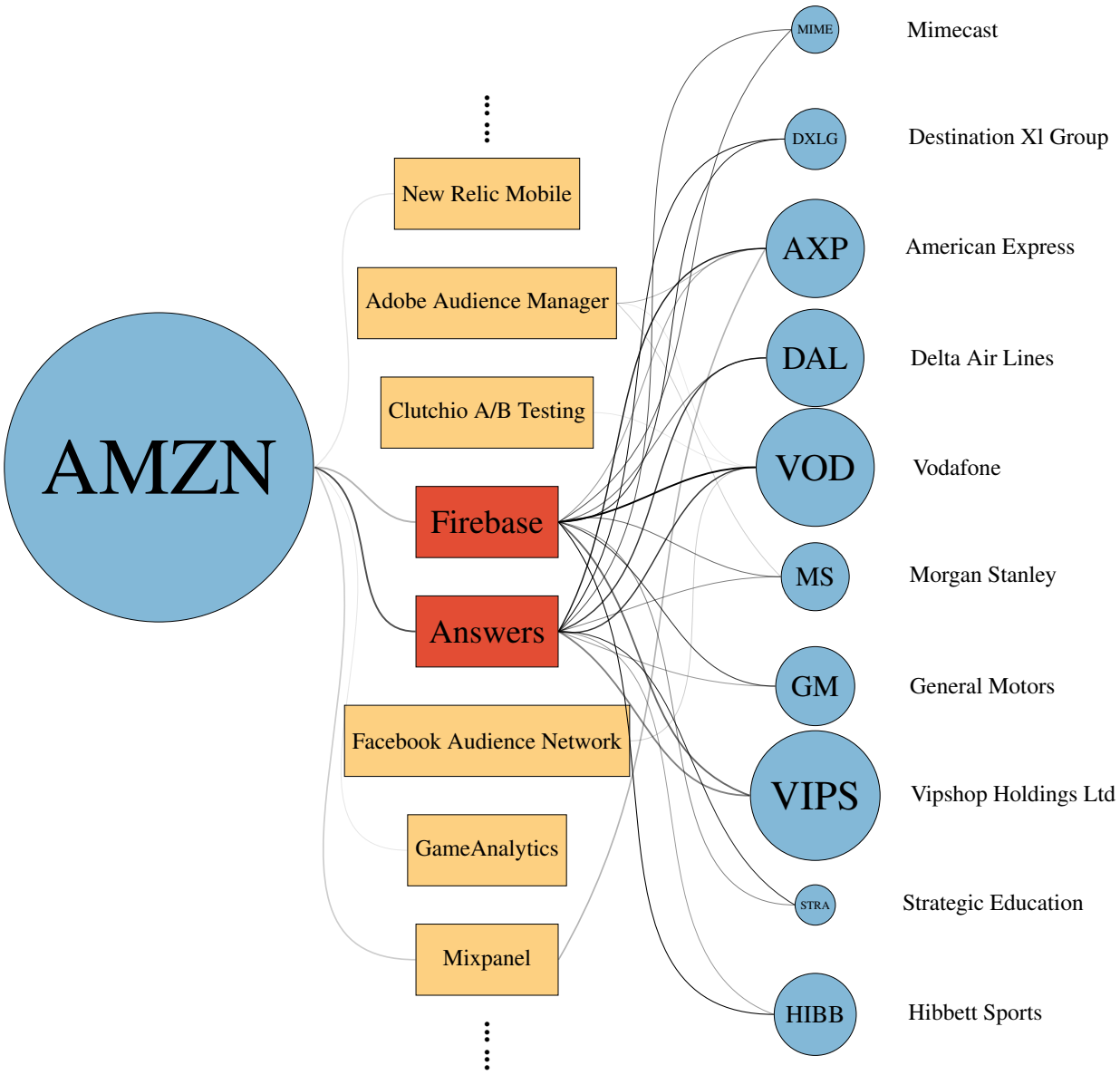


Panel B. Sales/Assets



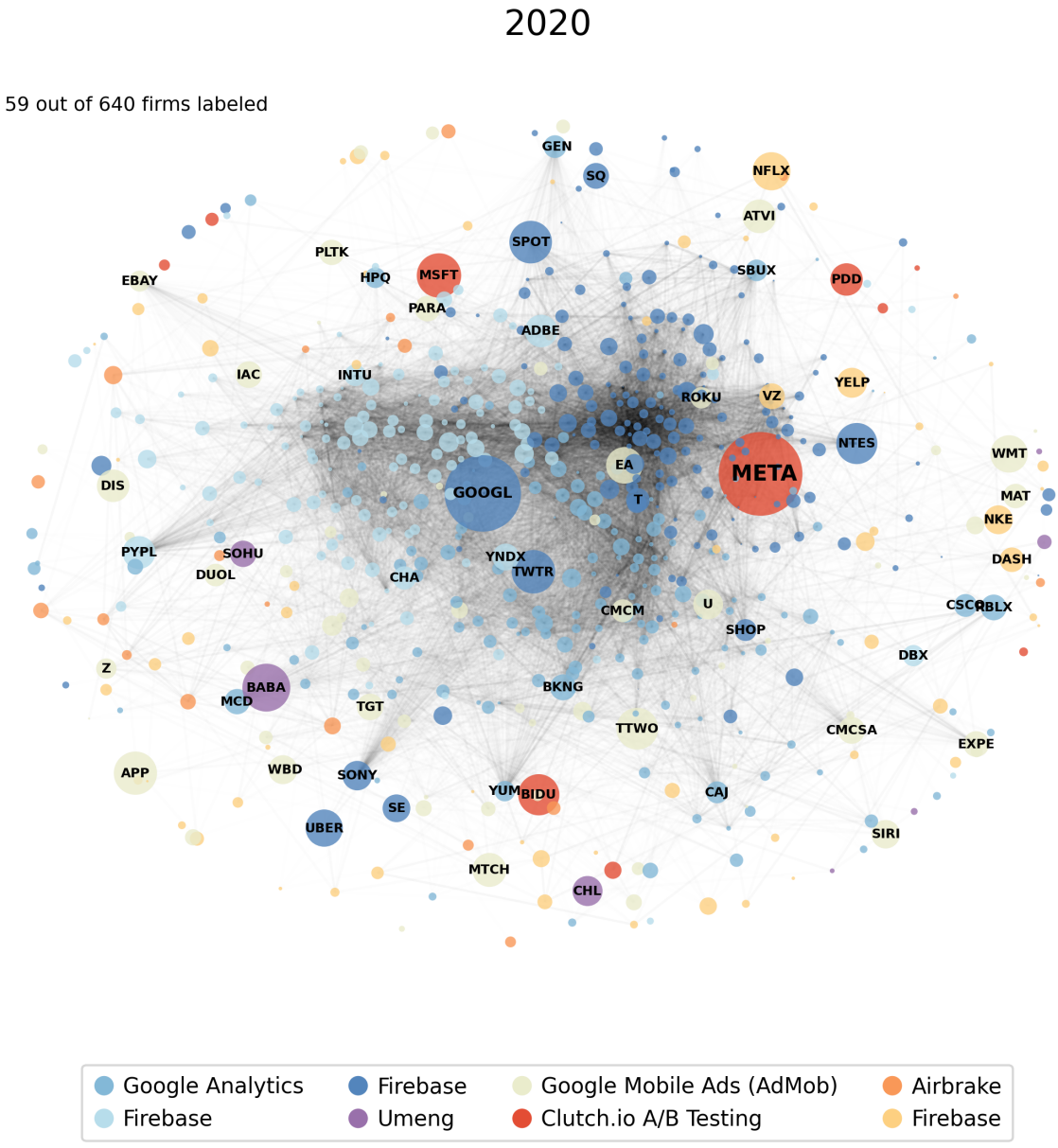
NOTE.—Figure 2 compares firms in the data network ("App Sample") with those in the broader Compustat universe ("Full Sample"). We focus on firm size, proxied by log(assets) (Panel A) and the sales/assets ratio (Panel B).

FIGURE 3: **Amazon's Data Peers**



NOTE.— Figure 3 illustrates Amazon and the 10 firms most connected to it in the data space. Firms are depicted as blue circles, with the circle size corresponding to the firm's Monthly Active Users (MAU). Data-related SDKs are shown as red or yellow rectangles. Lines connecting firms to SDKs indicate that the firm is utilizing the SDK, with the thickness of the line representing the relative importance of the SDK to the firm, measured by the proportion of the firm's MAU linked to the SDK.

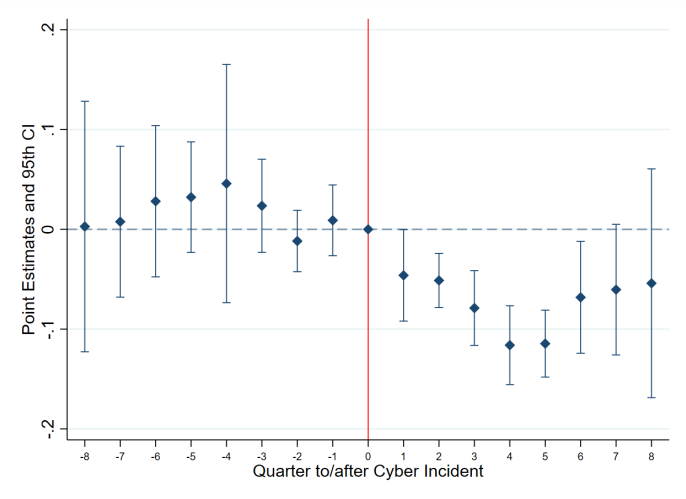
FIGURE 4: **The Data-Sharing Network**



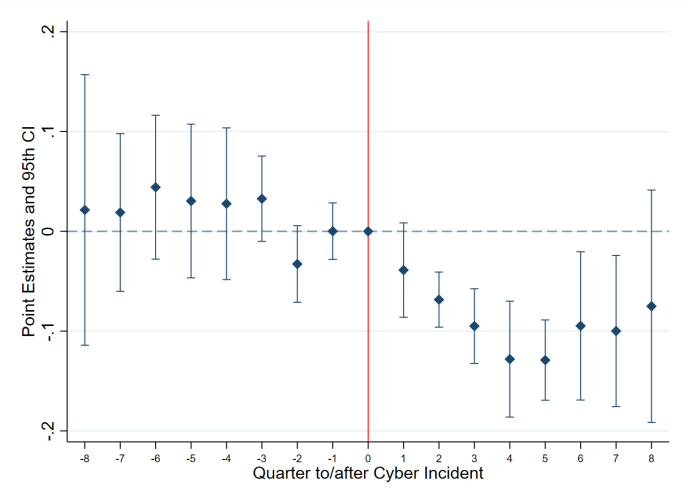
NOTE.—Figure 4 visualizes the network of firms connected in the data space using the Fruchterman-Reingold Algorithm and average data connectedness in 2020. Each node represents a firm, with the firm's ticker displayed, and the size of the node corresponds to the firm's size, proxied by the square root of total MAU in 2020. Firms connected by edges are those with positive data connectedness. For readability, we only include firm pairs with data connectedness greater than 0.7, which includes 640 unique firms and around 6.2% of all firm pairs. We label the firms that have more than 4.7 million MAU in an average quarter in 2020. Firms situated at the center typically have more highly-connected peers. Firms in different clusters, as identified by the Fruchterman-Reingold Algorithm, are distinguished by different colors. We identify and label the most popular SDK within each cluster of firms ex post.

FIGURE 5: **Cross-firm Spillover of Major Cyber Events on App Performance**

Panel A. ln(downloads)



Panel B. ln(DAU)



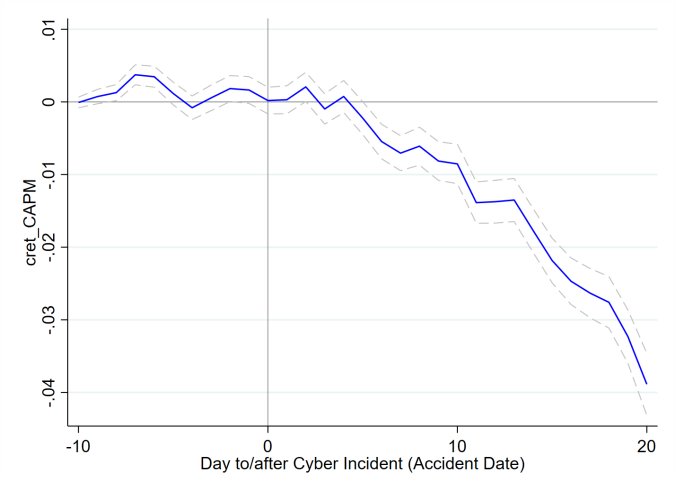
NOTE.— Figure 5 shows the cross-firm spillover effects of major cyber events on app performance. Major cyber events are defined as those that result in the exposure of over 10 million records. The dynamic DiD coefficients are obtained from estimating the dynamic version of Equation (3). For each firm k involved in a major cyber event, we define a peer firm i's exposure to the event as:

$$\text{Exposure}_{ik} = \frac{\sum_P \rho_{ik}^{\text{data},P} DAU_k^P}{\sum_P \sum_j \rho_{ij}^{\text{data},P} DAU_j^P}$$

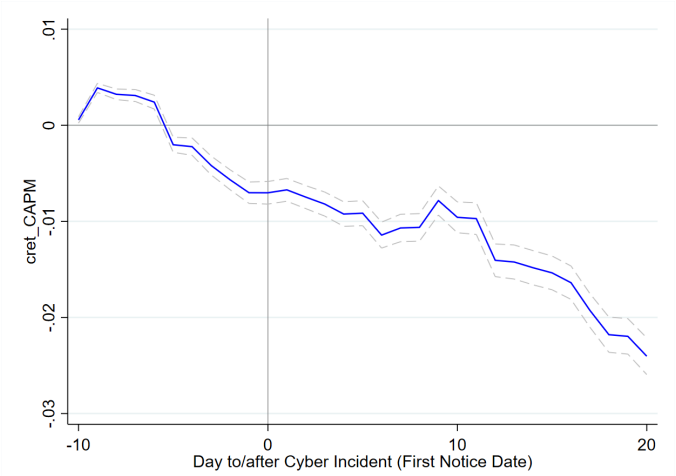
where j represents any other firm connected to firm i within the data space, and P represents platforms, taking values from {iOS, Android}. A firm k is considered an important peer if $\text{Exposure}_{ik} > 0.01$, corresponding to the 75th percentile of the exposure distribution. Firms with $\text{Exposure}_{ik} <= 0.01$ are considered as control firms. We include the following firm-level controls: firm size (log of assets), long-term debt to assets, and tangible assets to total assets. Additionally, we control for firm×event fixed effects and event-specific relative quarter fixed effects. Standard errors are double clustered by event and firm.

FIGURE 6: **Event Study of Cross-firm Spillover of Major Cyber Incidents**

Panel A. Event Date: Incident Date



Panel B. Event Date: Notice Date



NOTE.— Figure 6 presents event studies examining the cross-firm spillover effects of major cyber events. Panels A and B use the incident dates and notice dates, respectively, as the event dates. Both subfigures display the cumulative abnormal returns for peer stocks with high exposure to the event, using CAPM as the benchmark model. Major cyber events are defined as those that result in the exposure of over 10 million records. For each firm k involved in a major cyber event, we define a peer firm i's exposure to the event as:

$$\text{Exposure}_{ik} = \frac{\sum_P \rho_{ik}^{\text{data},P} DAU_k^P}{\sum_P \sum_j \rho_{ij}^{\text{data},P} DAU_j^P}$$

where j represents any other firm connected to firm i within the data space, and P represents platforms, taking values from {iOS, Android}. A firm k is considered an important peer if $\text{Exposure}_{ik} > 0.01$, corresponding to the 75th percentile of the exposure distribution.

FIGURE 7: **A Summary of Model Structure**

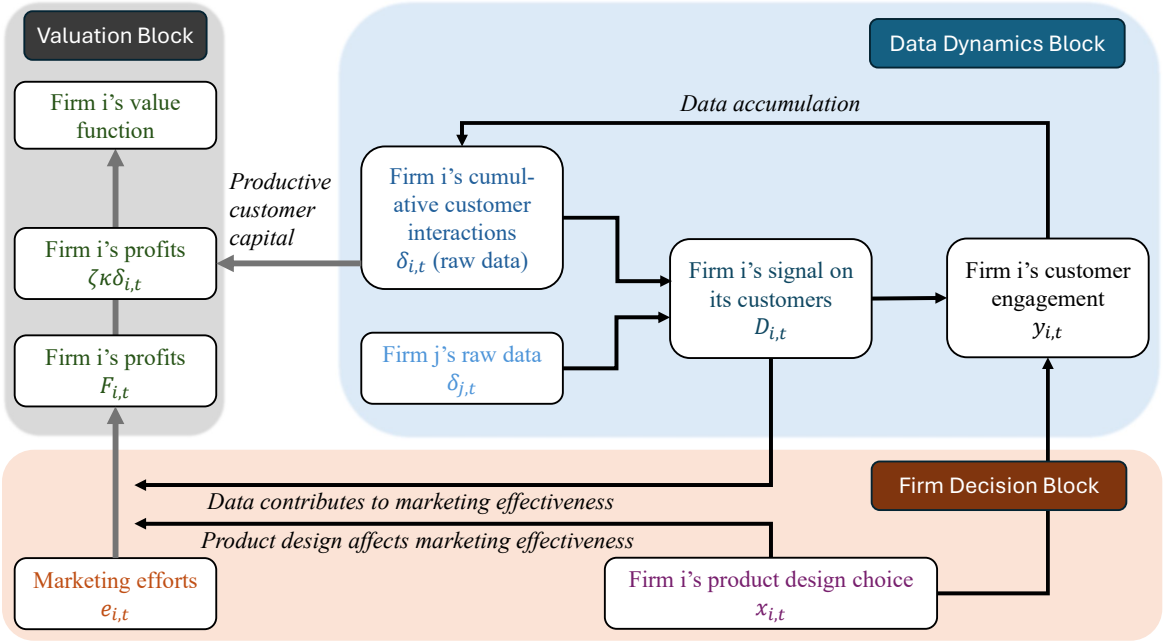
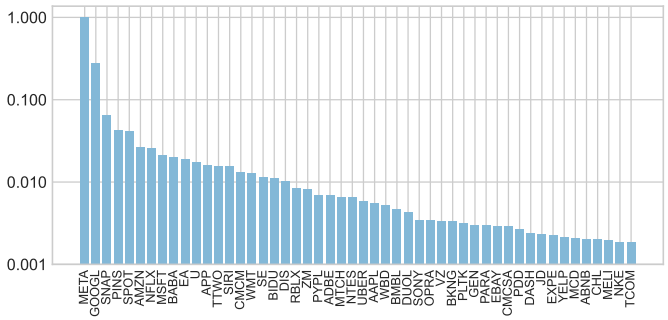
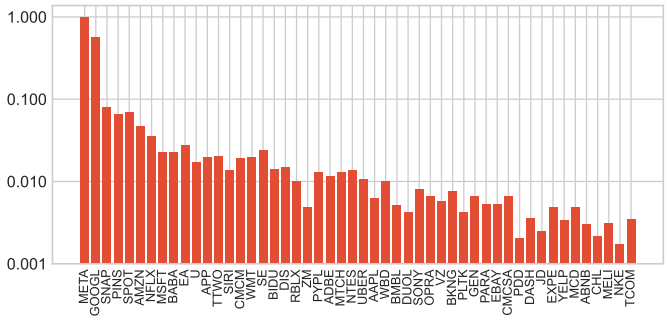


FIGURE 8: **Systemically Important Firms**

A. Top 50 DAU firms



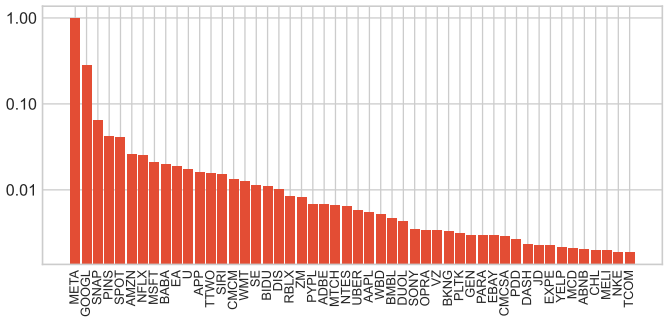
B. Valuation Contribution



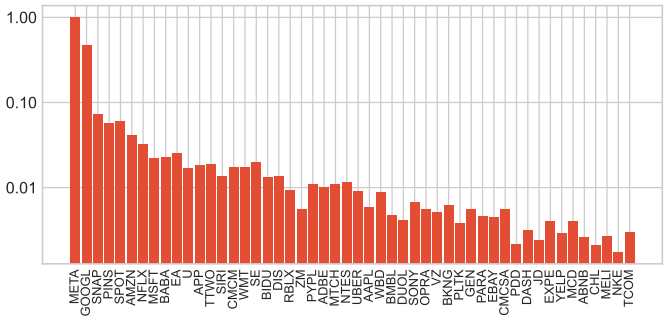
NOTE.—Figure 8 Panel A displays the top 50 firms ranked by their average daily active users (DAU) within the sample. The y-axis (in logs scale) represents the average DAU, with values scaled so that the maximum is normalized to 1. Firms are ordered by their average DAU from left to right. Panel B shows each firm's contribution to the total valuation of the network. Each bar corresponds to the element $v_{i,0} + \frac{\zeta\kappa}{\bar{\rho}} \left\{ \mathbf{1}^\top \left(\mathbf{I} - \frac{\beta}{\bar{\rho}} \Gamma \right)^{-1} \right\}_{.i} \delta_i$, representing each firm's impact on valuation. Panel C illustrates each firm's contribution to the total variance of the network. Each bar corresponds to the element $\left\{ \mathbf{1}^\top \left(\mathbf{I} - \frac{\beta}{\bar{\rho}} \Gamma \right)^{-1} \right\}_{.i} \sigma_i \delta_i$, representing each firm's impact on variance.

FIGURE 9: **Valuation Contribution under Different Degrees of Network Propagation**

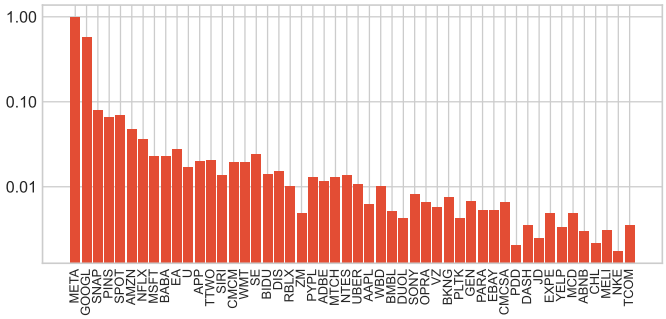
A. Valuation Contribution, $K = 0$



B. Valuation Contribution, $K = 3$

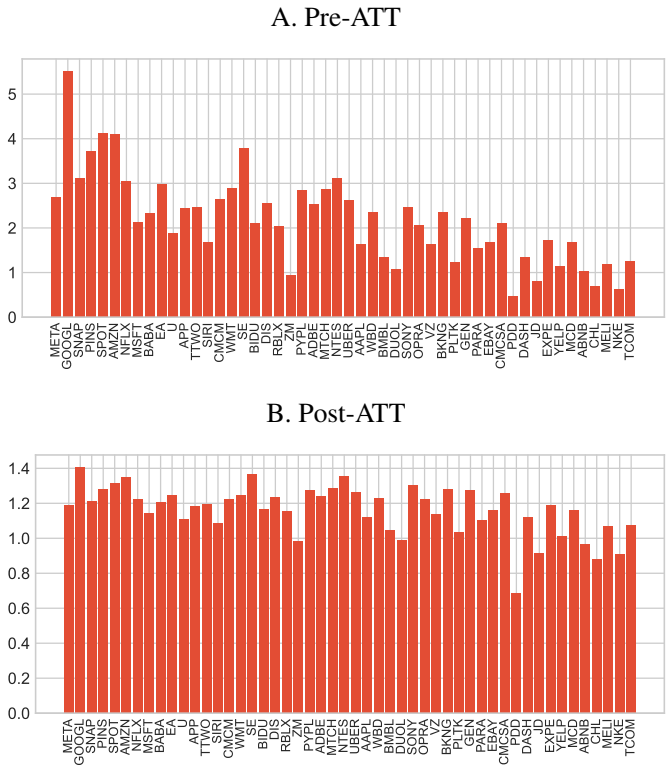


C. Valuation Contribution, $K = \infty$



NOTE.—Figure 9 illustrates the distribution of valuation contribution at different levels of network propagation. The y-axis (in logs scale) represents each firm's contribution to the total valuation of the network, with values scaled so that the maximum is normalized to 1. We display the top 50 firms ranked by their average daily active users (DAU) within the sample. Panel A shows the distribution when $K = 0$ (no network propagation). Panel B represents the case with $K = 3$ (shocks propagate three times through the network). Panel C corresponds to the case with $K = \infty$, when all network effects are accounted for.

FIGURE 10: **Firm Valuation vs. Firm Contribution to Aggregate Valuation**



NOTE.—Figure 10 compares the distributions of firms' valuation contributions to the total network, normalized by their individual valuations, before and after the ATT policy shock. Specifically, we compute the ratio of each firm's contribution to the total network valuation relative to its own valuation. We display the top 50 firms ranked by their average daily active users (DAU) within the sample. Panel A presents the distribution of this ratio prior to the introduction of the ATT policy, and Panel B shows the corresponding distribution after the policy's implementation.

TABLE 1: **Summary Statistics**

Panel A. Pair-wise variables

	mean	sd	p10	p25	p50	p75	p90	count
Pairwise connections								
data connectedness	0.170	0.20	0.00	0.00	0.10	0.29	0.46	1,401,082
mobile user (0/1)	0.036	0.19	0.00	0.00	0.00	0.00	0.00	1,401,082
app category	0.157	0.24	0.00	0.00	0.00	0.27	0.52	1,401,082
product horizontal	0.015	0.03	0.00	0.00	0.00	0.02	0.05	1,401,082
product vertical	0.003	0.00	0.00	0.00	0.00	0.00	0.01	1,401,082
technology	0.014	0.08	0.00	0.00	0.00	0.00	0.00	1,401,082
supply chain (0/1)	0.008	0.09	0.00	0.00	0.00	0.00	0.00	1,401,082
common analyst (0/1)	0.063	0.24	0.00	0.00	0.00	0.00	0.00	1,401,082
geography	0.300	0.43	0.00	0.00	0.00	0.90	0.99	1,401,082
Performance comovement								
downloads corr.	0.042	0.51	-0.67	-0.38	0.06	0.47	0.73	1,401,082
DAU corr.	0.073	0.57	-0.73	-0.42	0.11	0.59	0.81	1,395,990
earnings growth corr.	0.002	0.38	-0.49	-0.22	0.00	0.23	0.50	1,372,544
sales/assets corr.	0.095	0.46	-0.56	-0.25	0.12	0.46	0.71	1,223,256
Return comovement								
raw return	0.248	0.33	-0.21	0.02	0.27	0.50	0.67	5,720,040
return - CAPM	0.034	0.33	-0.41	-0.21	0.03	0.27	0.47	5,617,442
return - DGTW	0.008	0.33	-0.42	-0.23	0.01	0.24	0.44	4,967,560

Panel B. Firm-level variables

	mean	sd	p10	p25	p50	p75	p90	count
Δ payment SDK	0.008	0.28	0.00	0.00	0.00	0.00	0.00	20,344
Δ security SDK	0.005	0.14	0.00	0.00	0.00	0.00	0.00	20,344
Δ customer support SDK	0.001	0.08	0.00	0.00	0.00	0.00	0.00	20,344
Δ review & feedback SDK	0.002	0.07	0.00	0.00	0.00	0.00	0.00	20,344
L1.payment SDK (peers)	4.391	1.44	2.86	4.40	4.90	5.17	5.34	20,344
L1.security SDK (peers)	4.427	1.45	2.90	4.45	4.95	5.21	5.39	20,344
L1.customer support SDK (peers)	3.846	1.31	2.19	3.75	4.28	4.60	4.81	20,344
L1.review & feedback SDK (peers)	4.129	1.37	2.58	4.10	4.60	4.89	5.06	20,344
L1.payment SDK	2.023	2.06	0.00	0.00	2.00	4.00	5.00	20,344
L1.security SDK	0.827	0.84	0.00	0.00	1.00	1.00	2.00	20,344
L1.customer support SDK	0.139	0.39	0.00	0.00	0.00	0.00	1.00	20,344
L1.review & feedback SDK	0.333	0.54	0.00	0.00	0.00	1.00	1.00	20,344
L1.size	22.875	2.10	20.24	21.39	22.80	24.24	25.73	19,914
L1.long-term debt/assets	0.262	0.23	0.01	0.08	0.22	0.38	0.56	19,732
L1.tangibles/assets	0.203	0.22	0.01	0.04	0.11	0.30	0.57	19,457
L1.cash/assets	0.186	0.19	0.02	0.05	0.12	0.26	0.48	19,904

NOTE.—Table 1 reports the summary statistics on key variables. Panel A lists the all variables constructed at firm-pair level. For each firm pair, the comovement of app performance and financial performance is calculated separately for the periods before and after the introduction of ATT (2021Q2); return comovement is calculated as the correlation between their monthly returns over rolling 12-month windows, relative to the introduction of ATT in April 2021. The data on app performance, financial performance, and returns spans from 2014 September to 2023 June. Panel B includes variables constructed at the firm level, all at quarterly frequency.

TABLE 2: **Comovement in App Performance**

	downloads		DAU	
	(1)	(2)	(3)	(4)
data connectedness	0.025***	0.024***	0.025***	0.023***
	(7.58)	(7.31)	(6.71)	(6.25)
ATT × data connectedness	-0.025***	-0.025***	-0.025***	-0.024***
	(-6.90)	(-6.96)	(-6.31)	(-6.21)
mobile user (0/1)		0.001		0.023***
		(0.86)		(7.47)
app category		0.008***		0.010***
		(4.91)		(5.69)
product horizontal		0.009***		0.010***
		(4.63)		(5.83)
product vertical		0.001		-0.002
		(0.36)		(-0.96)
supply chain (0/1)		-0.000		-0.001
		(-0.61)		(-1.63)
technology		-0.001		0.000
		(-1.44)		(0.54)
common analyst (0/1)		0.005***		0.005***
		(4.99)		(4.11)
geography		0.003		0.002
		(1.57)		(0.93)
ATT × mobile user (0/1)		0.010***		-0.004
		(4.02)		(-1.02)
ATT × app category		-0.000		-0.003
		(-0.04)		(-0.97)
ATT × product horizontal		-0.001		-0.005*
		(-0.31)		(-1.92)
ATT × product vertical		0.003		0.011*
		(0.72)		(1.89)
ATT × supply chain (0/1)		0.000		0.001
		(0.22)		(0.72)
ATT × technology		0.000		-0.002
		(0.40)		(-1.33)
ATT × common analyst (0/1)		-0.002		-0.001
		(-1.41)		(-0.82)
ATT × geography		0.002		0.000
		(0.78)		(0.12)
Firm i#ATT FE	Y	Y	Y	Y
Firm j#ATT FE	Y	Y	Y	Y
Observations	1,401,082	1,401,082	1,399,426	1,399,426
R-sq	0.066	0.067	0.113	0.114

NOTE.—Table 2 shows the relationship between firm data connectedness and the comovement of app performance. Each observation represents a firm pair at a specific point in time. For each firm pair, the comovement of app performance is measured as the correlation between their quarterly log(downloads) and log(DAU), calculated separately for the periods before and after the introduction of ATT (2021Q2). The app performance data spans from 2014Q3 to 2023Q2. In even-numbered columns, we include controls for a comprehensive set of pair-wise firm connections, as well as interaction terms between these connections and the ATT indicator, which equals one for periods after 2021Q2. We include firm-by-time fixed effects (θ_{it} and ι_{jt}) and double cluster standard errors by firm-i and firm-j. Time is defined relative to ATT. Standard errors are double clustered by firm i and firm i. t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

TABLE 3: **Comovement in Firm Financial Performance**

	earnings growth		sales/assets	
	(1)	(2)	(3)	(4)
data connectedness	0.002***	0.002**	0.005***	0.004***
	(2.82)	(2.33)	(3.64)	(2.79)
ATT × data connectedness	-0.003***	-0.003***	-0.003**	-0.003**
	(-3.16)	(-2.88)	(-2.08)	(-2.07)
mobile user (0/1)		0.000		0.000
		(0.20)		(0.50)
app category		0.002**		0.005***
		(2.00)		(4.36)
product horizontal		0.005***		0.016***
		(3.98)		(9.25)
product vertical		0.002		0.007***
		(1.12)		(3.37)
supply chain (0/1)		-0.000		0.001
		(-1.16)		(1.08)
technology		-0.000		-0.001*
		(-0.61)		(-1.69)
common analyst (0/1)		-0.001		0.005***
		(-1.08)		(6.64)
geography		-0.002***		0.009***
		(-3.98)		(4.24)
ATT × mobile user (0/1)		0.000		0.003***
		(0.31)		(4.74)
ATT × app category		-0.002**		0.007***
		(-2.05)		(3.17)
ATT × product horizontal		-0.002		-0.006***
		(-1.25)		(-2.73)
ATT × product vertical		-0.004*		0.001
		(-1.80)		(0.16)
ATT × supply chain (0/1)		0.000		0.000
		(0.21)		(0.33)
ATT × technology		0.000		0.001
		(0.52)		(1.13)
ATT × common analyst (0/1)		0.001		-0.003***
		(0.81)		(-2.75)
ATT × geography		0.001		-0.005*
		(0.70)		(-1.73)
Firm i#ATT FE	Y	Y	Y	Y
Firm i#ATT FE	Y	Y	Y	Y
Observations	1,379,592	1,379,592	1,231,716	1,231,716
R-sq	0.005	0.005	0.184	0.186

NOTE.—Table 3 shows the relationship between firm data connectedness and the comovement of financial performance. Each observation represents a firm pair at a specific point in time. For each firm pair, the comovement of financial performance is measured as the correlation between their quarterly earnings growth and asset turnover (sales/assets), calculated separately for the periods before and after the introduction of ATT (2021Q2). The data on firm's financial performance spans from 2014Q3 to 2023Q2. In even-numbered columns, we include controls for a comprehensive set of pair-wise firm connections, as well as interaction terms between these connections and the ATT indicator, which equals one for periods after 2021Q2. We include firm-by-time fixed effects (θ_{it} and ι_{jt}) and double cluster standard errors by firm-i and firm-j. Time is defined relative to ATT. t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

TABLE 4: **Comovement in Stock Returns**

	raw return		return-CAPM		return-DGTW	
	(1)	(2)	(3)	(4)	(5)	(6)
data connectedness	0.004***	0.002***	0.005***	0.003***	0.003***	0.002***
	(6.34)	(3.87)	(6.57)	(4.56)	(6.00)	(3.71)
ATT × data connectedness	-0.002***	-0.002***	-0.003***	-0.003***	-0.001**	-0.002**
	(-2.69)	(-2.74)	(-3.38)	(-3.45)	(-2.19)	(-2.33)
mobile user (0/1)		0.001***		0.002***		0.001*
		(4.47)		(4.18)		(1.78)
app category		0.008***		0.009***		0.004***
		(8.57)		(7.87)		(5.22)
product horizontal		0.022***		0.030***		0.017***
		(22.21)		(21.48)		(19.48)
product vertical		0.002*		0.001		0.007***
		(1.70)		(0.98)		(4.26)
supply chain (0/1)		0.000		0.000		0.000
		(0.60)		(0.88)		(1.32)
technology		0.001***		0.002***		0.001
		(3.54)		(3.98)		(1.45)
common analyst (0/1)		0.009***		0.012***		0.009***
		(17.06)		(16.75)		(17.31)
geography		0.004***		0.004***		0.003***
		(4.52)		(3.91)		(3.70)
ATT × mobile user (0/1)		-0.001***		-0.002***		-0.000
		(-4.55)		(-4.72)		(-0.98)
ATT × app category		-0.001		-0.002		-0.001
		(-0.64)		(-1.50)		(-1.08)
ATT × product horizontal		0.001		-0.001		-0.001
		(0.53)		(-0.90)		(-1.22)
ATT × product vertical		0.001		0.003		-0.002
		(0.63)		(1.62)		(-1.29)
ATT × supply chain (0/1)		0.000		0.000		0.000
		(0.39)		(1.06)		(1.13)
ATT × technology		0.000		-0.000		0.000
		(0.62)		(-0.05)		(0.27)
ATT × common analyst (0/1)		0.001		0.002*		0.002***
		(1.29)		(1.87)		(2.66)
ATT × geography		-0.002		-0.002		0.000
		(-1.55)		(-1.25)		(0.19)
Firm i#Time FE	Y	Y	Y	Y	Y	Y
Firm j#Time FE	Y	Y	Y	Y	Y	Y
Observations	5,720,040	5,720,040	5,617,442	5,617,442	4,967,560	4,967,560
R-sq	0.442	0.450	0.097	0.110	0.022	0.028

NOTE.—Table 4 shows the relationship between firm data connectedness and return comovement. Each observation represents a firm pair at a specific point in time. For each firm pair, return comovement is calculated as the correlation between their monthly returns over rolling 12-month windows, relative to the introduction of ATT in April 2021. The return data spans from 2014 September to 2023 June, and we examine three types of returns: raw returns, abnormal returns based on CAPM, and DGTW-adjusted returns. In even-numbered columns, we include controls for a comprehensive set of firm-pair connections, along with interaction terms between these connections and the ATT indicator, which equals one for periods after 2021Q2. We include firm-by-time fixed effects (θ_{it} and ι_{jt}) and double cluster standard errors by firm-i and firm-j. Time is defined relative to ATT. Standard errors are double clustered by firm i and firm i. Standard errors are double clustered by firm i and firm j. t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

TABLE 5: **Cyberattack Spillover Effects on App Performance**

	log(downloads)		log(DAU)	
	(1)	(2)	(3)	(4)
cyber event × high exposure	−0.083**	−0.068**	−0.096***	−0.077**
	(−2.93)	(−2.31)	(−3.45)	(−2.76)
cyber event × mobile user		−0.162		2.512
		(−0.02)		(0.31)
cyber event × app category		−0.101		−0.160*
		(−1.38)		(−1.88)
cyber event × product horizontal		−0.050		−0.124
		(−0.11)		(−0.23)
cyber event × product vertical		8.174		8.415
		(0.83)		(0.68)
cyber event × technology		−0.129		−0.154
		(−1.49)		(−1.47)
cyber event × supply chain (0/1)		−0.058		−0.092
		(−1.05)		(−1.22)
cyber event × common analyst (0/1)		0.010		0.037
		(0.23)		(0.61)
cyber event × geography		−0.012		0.001
		(−0.21)		(0.01)
Firm controls	Y	Y	Y	Y
Firm#Event FE	Y	Y	Y	Y
Event-specific relative quarter FE	Y	Y	Y	Y
Observations	83,750	83,750	83,750	83,750
R-sq	0.954	0.954	0.955	0.955

NOTE.—Table 5 presents the cross-firm spillover effects of major cyber events using a stacked difference-in-differences (DiD) specification in 16-month event windows. Major cyber events are defined as those resulting in the exposure of over 10 million records. A comprehensive list of these events and their summaries can be found in Appendix Table B.1. For each firm k involved in a major cyber event, we define a peer firm i's exposure to the event as:

$$\text{Exposure}_{ik} = \frac{\sum_P \rho_{ik}^{\text{data},P} DAU_k^P}{\sum_P \sum_j \rho_{ij}^{\text{data},P} DAU_j^P}$$

where j represents any other firm connected to firm i within the data space, and P represents platforms, taking values from {iOS, Android}. A firm k is considered an important peer if $\text{Exposure}_{ik} > 0.01$, corresponding to the 75th percentile of the exposure distribution. Firms with $\text{Exposure}_{ik} <= 0.01$ are considered as control firms. Each regression includes the following firm-level controls: firm size (log of assets), long-term debt to assets, and tangible assets to total assets. Additionally, we control for firm×event fixed effects and event-specific relative quarter fixed effects. Standard errors are double clustered by event and firm. t-statistics are provided in parentheses. Statistical significance at the 1%, 5%, and 10% levels is denoted by ***, **, and *, respectively.

TABLE 6: **Model Calibration and Model-Implied Comovements**

A. Moments			
		Data	Model
Regression coefficient of			
– App performance on data network		0.023	0.038
– Financial performance on data network		0.004	0.003
– Return on data network		0.003	0.005
DID coefficient of			
– App performance on data network		-0.024	-0.034
– Financial performance on data network		-0.003	-0.003
– Return on data network		-0.003	-0.004
B. Parameters			
Description	Source	Notation	Quarterly Value
Strength of network propagation	Moments	β	0.080
Size of the ATT shock	Moments	ξ	0.700
Volatility of customer activities	Data	$\{\sigma_{\delta,i}\}_{i=1}^{N}$	—
Data network matrix	Data	$\{\gamma_{ij}\}_{i,j=1}^{N}$	—
Profits per customer	Data	ζ	18
Data depreciation rate	External	μ	-0.075
Discount rate	External	ρ	0.030
Proportion of paying customer	External	κ	1.000

NOTE.—Table 6 presents the calibration of the model. Panel A describes the set of moments that we target, and panel B presents the calibrated parameters. The target set of moments, shown in the upper panel, include the regression slopes of comovement between APP performance and stock returns on data network linkages, as well as the corresponding DID estimates that related to APP policy shocks. In the lower panel, we show the estimated parameters.

TABLE 7: **Product-Design Decisions: Monetization versus User Engagement**

Panel A. Payment

	Δ payment SDK	
	(1)	(2)
L1.payment SDK (peers)	0.007***	0.007***
	(5.60)	(5.25)
ATT × L1.payment SDK (peers)	−0.010***	−0.010***
	(−3.86)	(−3.80)
L1.payment SDK	−0.013***	−0.014***
	(−5.58)	(−5.75)
L1.size		0.006***
		(5.48)
L1.long-term debt/assets		0.003
		(0.31)
L1.tangibles/assets		0.013
		(1.03)
L1.cash/assets		0.033***
		(3.05)
Industry#Quarter FE	Y	Y
Firm controls	N	Y
Observations	20,344	19,274
R-sq	0.061	0.067

Panel B. User engagement

	Δ security SDK		Δ support SDK		Δ review SDK	
	(1)	(2)	(3)	(4)	(5)	(6)
L1.security SDK (peers)	0.004***	0.003***				
	(6.31)	(5.21)				
L1.customer support SDK (peers)			0.002***	0.001***		
			(4.32)	(4.27)		
L1.review & feedback SDK (peers)					0.001***	0.001***
					(5.05)	(4.13)
ATT × L1.security SDK (peers)	−0.003*	−0.002				
	(−1.70)	(−1.08)				
ATT × L1.customer support SDK (peers)			−0.002**	−0.002***		
			(−2.42)	(−2.78)		
ATT × L1.review & feedback SDK (peers)					−0.002***	−0.002***
					(−4.33)	(−3.80)
L1.security SDK	−0.015***	−0.016***				
	(−7.56)	(−7.52)				
L1.customer support SDK			−0.021***	−0.018***		
			(−5.40)	(−5.40)		
L1.review & feedback SDK					−0.008***	−0.008***
					(−5.79)	(−5.73)
Industry#Quarter FE	Y	Y	Y	Y	Y	Y
Firm controls	N	Y	N	Y	N	Y
Observations	20,344	19,274	20,344	19,274	20,344	19,274
R-sq	0.082	0.083	0.061	0.066	0.066	0.065

NOTE.—Table 7 shows cross-firm momentum in the investment in data accumulation. Panel A reports the results on payment SDK and Panel B on SDKs that are likely to improve user engagment. Standard errors are clustered by firm. t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

TABLE 8: **Product-Design Decisions: Model vs. Data**

	Payment	Security	Support	Model
Regression coefficient of				
– Change in firm's SDK on peers choices	0.007	0.004	0.001	0.006
DID coefficient of				
– Change in firm's SDK on peers choices	-0.010	-0.003	-0.002	-0.005

NOTE.—Table 8 presents the estimated regression coefficients from both the model and the data. The regression specification follows . A positive coefficient indicates that herding behavior in firms' production choices is influenced by data connectedness. Additionally, in both the model and the data, this herding behavior in firms' product design choices is significantly reduced after the introduction of the ATT policy. The parameters are estimated as in Table 6.

Internet Appendix to
"Data as a Networked Asset"

A Functionality SDKs

TABLE A.1: Description of Example Functionality SDKs

SDK Name	Category	# Installation [1]	Introduction
AliPay	Payment	23571	The cross-border app payment solution provides a convenient, safe, and reliable payment services to third-party applications. This payment solution is applicable to wireless devices (including mobiles and tablet computers) supported by Android or iOS system.
Stripe	Payment	8435	The Stripe SDK allows you to quickly build a payment flow in your app. We provide powerful and customizable UI elements that you can use out-of-the-box to collect your users' payment details.
Nimbus	Security	35884	The Nimbus SDK handles both requesting the ad and rendering the impression — all with a lightning-fast server-to-server connection — making it the easiest way to integrate with the Nimbus exchange. The SDK is customizable. You can choose to use the rendering function, the requesting function, or both.
Okta	Security	27663	Okta connects any person with any application on any device. It's an enterprise-grade, identity management service, built for the cloud, but compatible with many on-premises applications. With Okta, IT can manage any employee's access to any application or device.
Zendesk Support	Customer Support	4032	The SDK provides the following UIs for both Support and Guide to embed customer service features in an app: Help Center Overview - Lets the user access articles in your Zendesk Guide knowledge base and, optionally, submit a ticket. See Adding your help center; Help Center Article - Lets the user view a specific help center article. See Show a single article; Request - Lets the user view, update, and submit tickets to your customer service team. See Show a ticket screen; Request List - Lets the user view a list of their tickets. See Show the user's tickets.
Helpshift	Customer Support	2278	The Helpshift SDK allows your support team to provide in-app help in the form of searchable, native FAQs and direct, two-way messaging to end users.
Appirate	Reviews & Feedback	30384	Appirater is a class that you can drop into any iPhone app (iOS 4.0 or later) that will help remind your users to review your app on the App Store.
iRate	Reviews & Feedback	26894	iRate is a library to help you promote your iPhone and Mac App Store apps by prompting users to rate the app after using it for a few days.

B Major Cyber Events

TABLE B.1: List of Major Cyber Events

Company Name	Exposed Records	Date of Accident	Date of Notice	Case Type	Case Description
Baidu (China) Co., Ltd.	2 billion	13/05/2017	14/05/2017	Data – Malicious Breach	The DU Caller app, developed by Baidu's subsidiary, illegally stored users' personal data and secretly transferred contacts to its servers, which were hacked, exposing 2 billion phone numbers.
Marriott Int'l Inc	500 million	08/09/2018	19/11/2018	Data – Malicious Breach	On 8/9/2018, Marriott discovered an unauthorized attempt to access, encrypt, and remove data from its Starwood database. By 19/11/2018, Marriott believed data from up to 500 million guests had been compromised, including personal details for 327 million guests, with payment card information exposed for some.
Microsoft Corporation	250 million	28/12/2019	29/12/2019	Data – Unintentional Disclosure	Microsoft exposed call center data for nearly 250 million customers through several unsecured cloud servers, which was discovered by security researcher Bob Diachenko after the databases were indexed by the BinaryEdge search engine. The data spanned 14 years of Microsoft Customer Service and Support (CSS) records, which contained customer email and IP addresses, support agent emails, and internal notes. Microsoft secured the data by December 31, after being alerted on December 29.
Equifax Information Services of Puerto Rico Inc.	243 million	29/07/2017	12/09/2017	Privacy – Unauthorized Contact or Disclosure	On July 29, 2017, Equifax discovered a breach in its servers that exposed sensitive personal information, including the names, Social Security numbers, birth dates, and addresses of Michael W. Tomlin and Marilyn Tomlin. Equifax created a website for individuals to check if their data was compromised, with reports suggesting the breach affected over 100 million people. This incident resulted in Equifax violating the Fair Credit Reporting Act (FCRA).
Equifax Inc.	243 million	29/07/2017	20/09/2017	Phishing, Spoofing, Social Engineering	Software engineer Nick Sweeting created a fake version of Equifax's breach information site, equifaxsecurity2017.com, highlighting how easily the site could be impersonated. Several posts from Equifax's Twitter account mistakenly directed users to Sweeting's site, which received around 200,000 hits before being blacklisted by major browsers like Chrome, Firefox, and Safari. Equifax later deleted the incorrect links and apologized for the confusion.
Equifax Inc.	146 million	13/05/2017	30/07/2017	Data – Malicious Breach	In 2017, Equifax experienced a significant cybersecurity breach caused by criminals exploiting a vulnerability in the Apache Struts framework (CVE-2017-5638), affecting U.S., Canadian, and U.K. consumers. The attack affected occurred from mid-May to July 2017 and compromised names, Social Security numbers, birth dates, addresses, and in some cases, driver's license numbers. Equifax was notified of the vulnerability in March 2017 but failed to patch it until July 29, after detecting suspicious network activity. Initially, 145.5 million Americans were identified as affected, with an additional 2.4 million U.S. victims later identified whose names and partial driver's license information were stolen. Credit card details of 209,000 consumers and personal dispute documents of 182,000 were also accessed. In February 2020, U.S. authorities charged four Chinese military officers for the breach, alleging they sought Equifax's sensitive consumer data and trade secrets through the exploited vulnerability.
Capital One Financial Corp.	106 million	22/03/2019	19/07/2019	IT – Configuration/Implementation Errors	The breach was discovered on July 17, 2019, when a GitHub user alerted Capital One about a potential data theft, which the bank confirmed on July 19. Paige A. Thompson, an employee at a cloud computing company that provided data services to Capital One, was arrested for the breach after posting about it on GitHub. She exploited a misconfigured web application firewall to steal data from Capital One's servers. The breach impacted 106 million people, compromising transaction data, credit scores, payment history, balances, and for some, linked bank accounts and social security numbers.
Chex Systems Inc	100 million	24/09/2015	21/06/2016	Privacy – Unauthorized Contact or Disclosure	On September 24, 2015, Mission Bank sent Nicholas A. George a letter refusing to open a deposit account, citing information from a consumer report obtained from Chex Systems, Inc. (Chex). George later obtained his ChexSystems report on February 6, 2016, discovering that the Academy Bank trade line inaccurately reflected his liability for the account. This incorrect reporting harmed George by causing embarrassment, inconvenience, and annoyance. Due to its size and large consumer database, Chex's actions violated the Fair Credit Reporting Act (FCRA), harming the hundreds of millions of consumers for whom it holds banking history data.

Google LLC	53 million	07/11/2018	10/12/2018	IT – Configuration/Implementation Errors	On December 10, 2018, Google disclosed a second bug in the Google+ API that potentially exposed the private data of 52.5 million users. Discovered during internal tests, Google stated there was no evidence that third parties had exploited the bug. The issue, caused by a software update, affected Google+ APIs between November 7 and November 13, 2018, when it was fixed. As a result, Google moved the shutdown of Google+ for consumers from August 2019 to April 2019. The bug in the Google+ People API allowed apps to access profile data, including names, emails, and birthdays, which users had marked as private. More sensitive information, such as passwords and financial data, was not affected. Google has since notified impacted users.
T-Mobile US, Inc.	50 million	19/08/2021	07/10/2021	Data – Malicious Breach	Edward Mendez was a victim from SIM-swapping attacks on August 19 and September 12, 2021, with a loss of nearly $240,000 in cryptocurrency. The employee who granted the hacker access had bypassed the 'text-message notification' protocol that notifies all other members under the same account when there is a change to an account. The hackers also disabled two-factor authentication and accessed Mendez's Coinbase account, changing his password and deleting related emails. The breach exposed sensitive information, including security numbers, phone numbers, addresses, and driver's license details. The Kansas attorney general reported that over 335,000 Kansas residents could be affected by the T-Mobile data breach.
T-Mobile US Inc	50 million	16/08/2021	18/08/2021	Data – Malicious Breach	The breach was detected after the attacker reported the incident to Motherboard. On August 16, T-Mobile confirmed the breach, which affected 7.8 million current postpaid customers and over 40 million records of former or prospective customers who applied for credit. The company claimed that the stolen data included personal information such as names, birthdates, Social Security numbers, and driver's license/ID numbers, but not bank, payment data, or passwords. Additionally, the names, phone numbers, and account PINs of around 850,000 prepaid users were exposed. T-Mobile quickly shut down the access point used in the attack. As a consequence, the company is offering two years of free identity theft protection via McAfee and advising postpaid customers to change their PINs while also providing account takeover protection.
T-Mobile Usa, Inc.	50 million	17/08/2021	24/08/2021	Data – Malicious Breach	On August 17, 2021, T-Mobile discovered that a bad actor had illegally accessed unencrypted personal information, which included names, driver's license numbers, phone numbers, addresses, government identification numbers, Social Security numbers, dates of birth, and T-Mobile account PINs.
Chegg Inc	40 million	29/04/2018	19/09/2018	Data – Malicious Breach	Chegg, Inc., a US-based education technology company, plans to reset passwords for over 40 million users after discovering a security breach that dates back to April 29, 2018. The breach was detected on September 19, 2018, and involved unauthorized access to a company database containing user data for chegg.com and related brands, such as EasyBib. Hackers may have accessed user information, including names, email addresses, shipping addresses, usernames, and hashed passwords, although Chegg did not specify the hashing algorithm used. Social Security numbers and financial data were not compromised. The breach caused Chegg's stock price to drop by 10 percent.
T-Mobile US Inc	37 million	25/11/2022	05/01/2023	Data – Malicious Breach	On January 19, 2023, T-Mobile reported a cyberattack that exposed data from approximately 37 million postpaid and prepaid customer accounts. The breach was detected on January 5, 2023, when T-Mobile identified unauthorized data access through a single Application Programming Interface (API). The API did not expose sensitive information such as payment card details, Social Security numbers, or passwords, but did allow access to customer data including names, billing addresses, emails, phone numbers, birth dates, account numbers, and plan details.
Taobao	21 million	14/10/2015	04/02/2016	Data – Malicious Breach	From October 14 to 16, 2015, a group of hackers attempted to access over 20 million active user accounts on Taobao, Alibaba Group's e-commerce platform, using rented space on Alibaba's AliCloud services. Of the 99 million accounts involved, 20.59 million had matching passwords. The hackers aimed to acquire these accounts for order manipulation and sale to scammers. However, the attack did not involve a direct breach of Taobao. Instead, hackers used account information from non-Taobao platforms to find matching credentials. The hack was stopped a month later by Chinese authorities after website admins detected suspicious activity on the platform.
Morgan Stanley	14 million	21/02/2020	10/07/2020	Data – Malicious Breach	In 2019, Morgan Stanley replaced certain computer servers in local branch offices that stored information on encrypted disks, which may have contained personal data. During an inventory, Morgan Stanley was unable to locate these encrypted disks, leading to a data breach. The incident, which occurred on February 21, 2020, compromised personally identifiable information, including Social Security numbers, affecting 14,256,250 individuals.

Twitter Inc	13 million	07/02/2020	07/02/2020	Data – Malicious Breach	On February 7, 2020, the official Facebook Twitter account was briefly taken over by the hacking group OurMine. The incident lasted less than 30 minutes, during which a tweet was sent to Facebook's 13.4 million followers, stating: "Hi, we are OurMine. Well, even Facebook is hackable but at least their security better than Twitter," and offering "security services" to improve account protection. The breach was not a result of compromised Facebook or Twitter systems, but rather due to a third-party marketing platform used to manage social media. A Twitter spokesperson confirmed the issue, stating the compromised accounts were quickly locked. Facebook later confirmed in a tweet that the issue had been resolved and access restored.
Blackbaud Inc	13 million	07/02/2020	01/05/2020	Data – Malicious Breach	In May 2020, Blackbaud, Inc. was targeted in a sophisticated ransomware attack. The breach, which began on February 7, 2020, and lasted intermittently until May 20, 2020, compromised backup files for clients using Blackbaud's Raiser's Edge/NXT system. While the hackers did not access encrypted credit card information, bank account details, Social Security numbers, or login credentials, they did obtain contact information, demographic data, and donation histories. Blackbaud paid an undisclosed ransom after evidence showed the stolen data was destroyed, and it is believed the compromised data was not misused or publicly shared. However, further investigation revealed that more unencrypted data, including bank account information and Social Security numbers, may have been accessed. As of September 2020, the Identity Theft Resource Center reported that 536 organizations and nearly 13 million people were impacted.
Quest Diagnostics Inc	12 million	01/08/2018	14/05/2019	Data – Malicious Breach	On June 3, 2019, Quest Diagnostics revealed that a data breach potentially exposed the personal, financial, and medical information of approximately 11.9 million patients. The breach occurred through a billing collections vendor, American Medical Collection Agency (AMCA), which provides services to Optum360, a Quest contractor. An unauthorized user had access to AMCA's system from August 1, 2018, to March 30, 2019. The compromised data included credit card numbers, bank account information, medical details, and Social Security numbers, though lab results were not exposed. As of May 31, 2019, AMCA estimated that 11.9 million Quest patients were affected. AMCA has yet to provide full details about the breach, and Quest has been unable to verify all of the information.
MGM Resorts International	11 million	07/07/2019	21/02/2020	Data – Malicious Breach	On or around July 7, 2019, an unauthorized individual accessed MGM Resorts International's computer network and stole customer data, which included personal information such as names, addresses, driver's license and passport numbers, military IDs, phone numbers, emails, and dates of birth. A subset of this data was initially shared on a closed internet forum but was later fully exposed on a hacking forum in February 2020, affecting over 10.6 million MGM guests. This breach left customers vulnerable to phishing attacks and SIM-swapping schemes. Despite the breach occurring seven months prior, MGM did not publicly disclose it until September 5, 2019, when it notified affected customers and government agencies, due to a belief that the data wouldn't be misused.
Laboratory Corp of America Holdings	10 million	01/08/2018	14/05/2019	Data – Malicious Breach	On May 14, 2019, Laboratory Corporation of America Holdings (LabCorp) was notified by its vendor, Retrieval-Masters Creditors Bureau, Inc., operating as American Medical Collection Agency (AMCA), of unauthorized activity on AMCA's web payment page. The breach occurred between August 1, 2018, and March 30, 2019. LabCorp immediately ceased sending collection requests to AMCA and halted pending requests. AMCA, which serves as an external collection agency for LabCorp and other healthcare companies, stored data for approximately 7.7 million LabCorp consumers. The compromised data included personal information such as names, addresses, dates of birth, and payment details (credit card or bank account information). No laboratory results, test orders, Social Security numbers, or insurance details were exposed. The breach impacted a total of 10,241,756 consumers.
Chipotle Mexican Grill Inc	10 million	24/03/2017	25/04/2017	Data – Malicious Breach	On April 25, 2017, Chipotle disclosed a data breach caused by credit card-stealing malware that infected the payment processing system in most of its 2,250 restaurants. The malware collected cardholder information, including names, card numbers, expiration dates, and verification codes, during transactions between March 24 and April 18, 2017. Chipotle has since removed the malware. The breach affected tens of millions of customers, including 1,798 New Jersey residents.

C Model Derivation and Proofs

C.1 Customer optimization

Consider a consumer in a data economy, whose problem is to maximize their utility from consuming firm's product. The utility function is quasi-linear, reflecting diminishing marginal utility from consumption of the good, while the cost of consumption is linear in the quantity consumed. Specifically,

$$U = \max_{q} \zeta \log(q) - pq, \tag{C.1}$$

where q is the quantity of the good consumed, p is the price per unit of good. ζ is a positive constant reflecting the consumer's preference for consuming the good. Taking FOC and solving for optimal quantity:

$$q = \frac{\zeta}{p} \tag{C.2}$$

The optimal expenditure per customer is

$$pq = \zeta \tag{C.3}$$

Therefore, a firm that has a paying customer base of δ can generate profits $\zeta\delta$.

C.2 Proof of Lemma 1

Given that the profits function is defined as

$$F_{i,t} = \max_{e_{i,t}} \ \zeta\omega^m_{i,t}(e_{i,t}) - C(D_{i,t}, x_{i,t})e_{i,t}, \tag{C.4}$$

FOC with respect to $e_{i,t}$ yields

$$\zeta\omega'(e^m_{i,t}) = C(D_{i,t}, x_{i,t}), \tag{C.5}$$

where e^m denotes the optimal level of e. Note that ω is an increasing and concave function in e with $w''(e) < 0$, and that $C_D < 0, C_x > 0$. Therefore, differentiating both sides of (C.5) with respect to $D_{i,t}$, we have

$$\zeta\omega''(e)\frac{\partial e^m_{i,t}}{\partial D_{i,t}} = C_D(D_{i,t}, x_{i,t}). \tag{C.6}$$

That is,

$$\frac{\partial e_{i,t}^m}{\partial D_{i,t}} = \frac{C_D(D_{i,t}, x_{i,t})}{\zeta \omega''(e)} > 0 \tag{C.7}$$

Similarly, differentiating both sides of (C.5) with respect to $x_{i,t}$, we have

$$\zeta \omega''(e) \frac{\partial e_{i,t}^m}{\partial x_{i,t}} = C_x(D_{i,t}, x_{i,t}) \tag{C.8}$$

That is,

$$\frac{\partial e_{i,t}^m}{\partial x_{i,t}} = \frac{C_x(D_{i,t}, x_{i,t})}{\zeta \omega''(e)} < 0 \tag{C.9}$$

This also implies that

$$\frac{\partial \omega_{i,t}^m}{\partial D_{i,t}} > 0, \quad \frac{\partial \omega_{i,t}^m}{\partial x_{i,t}} < 0. \tag{C.10}$$

And the associated profits

$$F_{i,t} = \zeta \omega_{i,t}^m(e_{i,t}^m) - C(D_{i,t}, x_{i,t}) e_{i,t}^m. \tag{C.11}$$

Taking derivative, using the envelop theorem

$$\frac{\partial F_{i,t}}{\partial x_{i,t}} = -\frac{\partial C(D_{i,t}, x_{i,t})}{\partial x_{i,t}} e_{i,t}^m < 0, \tag{C.12}$$

and

$$\frac{\partial F_{i,t}}{\partial D_{i,t}} = -\frac{\partial C(D_{i,t}, x_{i,t})}{\partial D_{i,t}} e_{i,t}^m > 0. \tag{C.13}$$

Since we have $C_{xD} < 0$, the cross-derivative of $F_{i,t}$ with respect to $x_{i,t}$ and $D_{i,t}$ satisfies

$$\frac{\partial^2 F_{i,t}}{\partial D_{i,t} x_{i,t}} = -\frac{\partial^2 C(D_{i,t}, x_{i,t})}{\partial D_{i,t} \partial x_{i,t}} e_{i,t}^m > 0. \tag{C.14}$$

C.3 Proof of Proposition 1

Note that the optimal choice of $x_{i,t}$ is characterizied by the HJB equation,

$$\rho V^i\left(\delta_{i,t}, \{\delta_{j,t}\}_{j\neq i}\right) = \max_{x_{i,t}} \zeta\kappa\delta_{i,t} + F\left(D_{i,t}, x_{i,t}\right) + V^i_{\delta_{i,t}}\left[\theta(\alpha D_{i,t} + x_{i,t}) + \mu_\delta \delta_{i,t}\right] + \frac{1}{2}V^i_{\delta_{i,t}\delta_{i,t}}\delta^2_{i,t}\sigma^2_{i,\delta}$$
$$+ \sum_{j\neq i}\left[V^i_{\delta_{j,t}}\left[\theta(\alpha D_{j,t} + x_{j,t}) + \mu_{j,\delta}\delta_{j,t}\right] + \frac{1}{2}V^i_{\delta_{j,t}\delta_{j,t}}\delta^2_{j,t}\sigma^2_{j,\delta}\right]. \tag{C.15}$$

Taking FOC with respect to $x_{i,t}$, we have

$$-F_x(D_{i,t}, x_{i,t}) = V_{\delta_{i,t}}\theta \tag{C.16}$$

C.4 Proof of Propositions 2

For firm's cash flow from user activities, we have

$$F_{i,t} = \zeta\log(e_{i,t}) - \frac{e_{i,t}}{\phi_0 D_{i,t} - \phi_1 x_{i,t}} \tag{C.17}$$

Taking FOC with respect to $e_{i,t}$, we obtain

$$e_{i,t} = \zeta(\phi_0 D_{i,t} - \phi_1 x_{i,t}) \tag{C.18}$$

substituting back, firm's profits from user activities are

$$F\left(\delta_{i,t}, \{\delta_{j,t}\}\right) = \zeta\log\left(\phi_0 D_{i,t} - \phi_1 x_{i,t}\right) + \zeta\log\zeta - \zeta \tag{C.19}$$

where

$$D_{i,t} = \sum_{j=1}^{N}\gamma_{ij}\delta_{j,t}. \tag{C.20}$$

Value function conjecture. To solve for firm's valuation, we conjecture that the firm's value function has the following functional form:

$$V\left(\delta_{i,t}, \{\delta_{j,t}\}\right) = v_{i,0} + v_i^\top\bar{\delta}_t = v_{i,0} + \sum_{j=1}^{N} v_{i,j}\delta_{j,t} \tag{C.21}$$

where $\bar{\delta}_t$ is the column vector of all firms' data stock, $\bar{\delta}_t = [\delta_{1,t}, ..., \delta_{N,t}]^\top$. Therefore, we obtain the following expressions for firm i's dependency on firm j's data $\forall j = 1, 2, ...N$:

$$V_{\delta_{j,t}}\left(\delta_{i,t}, \{\delta_{j,t}\}\right) = v_{i,j},$$
$$V_{\delta_{j,t}\delta_{j,t}}\left(\delta_{i,t}, \{\delta_{j,t}\}\right) = 0,$$

we substitute these into the HJB equation to obtain:

$$\rho V\left(\delta_{i,t}, \{\delta_{j,t}\}\right) dt = \max_{x_{i,t}} \left(F\left(\delta_{i,t}, \{\delta_{j,t}\}\right) + \zeta\kappa\delta_{i,t}\right)dt + v_{i,i}\left(\theta(x_{i,t} + \alpha D_{i,t}) + \mu_\delta\delta_{i,t}\right) dt$$
$$+ v_{i,j}\sum_{j\neq i}^{N}\left[\theta(x_{j,t} + \alpha D_{j,t}) + \mu_\delta\delta_{j,t}\right] dt\,. \tag{C.22}$$

FOC (11) becomes

$$\frac{\zeta\phi_1}{(\phi_0 D_{i,t} - \phi_1 x_{i,t})} = \theta v_{i,i}. \tag{C.23}$$

If $v_{i,i} > 0$, this ensures that $\phi_0 D_{i,t} - \phi_1 x_{i,t} > 0$ and also gives

$$x_{i,t} = \frac{\phi_0}{\phi_1} D_{i,t} - \frac{\zeta}{\theta v_{i,i}}. \tag{C.24}$$

That is

$$x_{i,t} = \frac{\phi_0}{\phi_1} D_{i,t} - \frac{\zeta}{\theta v_{i,i}} = \frac{\phi_0}{\phi_1}(\sum_{j=1}^{N}\gamma_{ij}\delta_{j,t}) - \frac{\zeta}{\theta v_{i,i}} \tag{C.25}$$

Also, we have

$$e_{i,t} = \frac{\zeta^2\phi_1}{\theta v_{i,i}} > 0 \tag{C.26}$$

The cash flow from user activities is given by

$$F_{i,t} = \zeta\log(\frac{\zeta\phi_1}{\theta v_{i,i}}) + \zeta\log\zeta - \zeta \tag{C.27}$$

Substitute FOC into HJB

$$\rho V = \zeta\kappa\delta_{i,t} + \zeta\log(\frac{\zeta\phi_1}{\theta v_{i,i}}) + \zeta\log\zeta - \zeta + \sum_{j=1}^{N} v_{i,j}\left[\theta x_{j,t} + \alpha\theta D_{j,t} + \mu_\delta\delta_{j,t}\right]$$

$$= \zeta\kappa\delta_{i,t} + \zeta\log(\frac{\zeta\phi_1}{\theta v_{i,i}}) + \zeta\log\zeta - \zeta + \sum_{j=1}^{N} v_{i,j}\left(\theta\frac{\phi_0}{\phi_1}D_{j,t} - \frac{\zeta}{v_{j,j}} + \alpha\theta D_{j,t} + \mu_\delta\delta_{j,t}\right) \quad \text{(C.28)}$$

where

$$D_{j,t} = \left(\sum_{k=1}^{N}\gamma_{jk}\delta_{k,t}\right) \quad \text{(C.29)}$$

After substitution and simplification:

$$\rho V_{i,t} = \left(\zeta\kappa\delta_{i,t} + \sum_{j=1}^{N} v_{i,j}\left[\theta(\frac{\phi_0}{\phi_1} + \alpha)\left(\sum_{k=1}^{N}\gamma_{jk}\delta_{k,t}\right) + \mu_\delta\delta_{j,t}\right]\right) + A_i \quad \text{(C.30)}$$

where

$$A_i = \left(\zeta\log(\frac{\zeta\phi_1}{\theta v_{i,i}}) + \zeta\log\zeta - \zeta + \sum_{j=1}^{N} v_{i,j}\left[-\frac{\zeta}{v_{j,j}}\right]\right) \quad \text{(C.31)}$$

This gives the constant term in the valuation

$$v_{i,0} = \frac{A_i}{\rho}. \quad \text{(C.32)}$$

There are two components in it. The first is the discounted value of all future cash flow from user activities. The second component is the present value of the change in data accumulation resulting from optimal product design.

Next, by comparing coefficients on LHS and RHS of N states $\delta_{j,t}$, we have N equations for N unknown $v_{i,j}, \quad \forall j = 1, 2, ..., N$. The valuation vector is

$$v_i = \begin{pmatrix} v_{i,1} \\ v_{i,2} \\ \vdots \\ v_{i,N} \end{pmatrix}$$

Γ is an $N \times N$ matrix of network linkages

$$\Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1N} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} & \cdots & \gamma_{2N} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} & \cdots & \gamma_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{N1} & \gamma_{N2} & \gamma_{N3} & \cdots & \gamma_{NN} \end{pmatrix}$$

For (C.30), the coefficients on δ_i should be the same on both sides, therefore we have for $v_{i,i}$:

$$\rho v_{i,i} = \zeta\kappa + \theta(\frac{\phi_0}{\phi_1} + \alpha) \left(\sum_{j=1}^{N} \gamma_{ji} v_{i,j} \right) + \mu_\delta v_{i,i}$$

and for the coefficient of $\delta_{j,t}$ for $j \neq i$ we have

$$\rho v_{i,j} = \theta(\frac{\phi_0}{\phi_1} + \alpha) \left(\sum_{k=1}^{N} \gamma_{kj} v_{i,k} \right) + \mu_\delta v_{i,j}$$

Define:

$$\beta = \theta(\frac{\phi_0}{\phi_1} + \alpha)$$

The above equations can be written as

$$(\rho - \mu_\delta) v_{i,i} = \zeta\kappa + \beta(\Gamma^\top v_i)_i \tag{C.33}$$

and

$$(\rho - \mu_\delta) v_{i,j} = \beta(\Gamma^\top v_i)_j$$

Let's define

$$\hat{\rho} = \rho - \mu_\delta$$

Rearranging the above equations in a matrix form, we get:

$$(I\hat{\rho} - \beta\Gamma^\top) v_i = \zeta\kappa \mathbf{e}_i$$

Solve for the valuation vector v_i, we obtain

$$v_i = (I\hat{\rho} - \beta\Gamma^\top)^{-1} \zeta\kappa \mathbf{e}_i \tag{C.34}$$

Therefore, the valuation vector is given by

$$\begin{aligned} v_i &= (I\hat{\rho} - \beta\Gamma^\top)^{-1}\zeta\kappa\mathbf{e}_i \\ v_i^\top &= \mathbf{e}_i^\top(I - \frac{\beta}{\hat{\rho}}\Gamma)^{-1}\frac{\zeta\kappa}{\hat{\rho}} \\ &= \frac{\zeta\kappa}{\hat{\rho}}\mathbf{e}_i^\top\left(I + \frac{\beta}{\hat{\rho}}\Gamma + (\frac{\beta}{\hat{\rho}})^2\Gamma^2 + (\frac{\beta}{\hat{\rho}})^3\Gamma^3 + ...\right). \end{aligned}$$

And all elements of the vector are positive. Therefore, the firm's valuation is

$$V_{i,t} = v_{i,0} + \frac{\zeta\kappa}{\hat{\rho}}\mathbf{e}_i^\top\left(I + \frac{\beta}{\hat{\rho}}\Gamma + (\frac{\beta}{\hat{\rho}})^2\Gamma^2 + (\frac{\beta}{\hat{\rho}})^3\Gamma^3 + ...\right)\bar{\delta}_t \tag{C.35}$$

$$= v_{i,0} + \eta\delta_i + \eta\mathbf{e}_i^\top\frac{\beta}{\hat{\rho}}(I + \frac{\beta}{\hat{\rho}}\Gamma + (\frac{\beta}{\rho})^2\Gamma^2 + ...)\Gamma\bar{\delta}_t \tag{C.36}$$

$$= v_{i,0} + \eta\delta_i + \eta\mathbf{e}_i^\top\frac{\beta}{\hat{\rho}}(I - \frac{\beta}{\hat{\rho}}\Gamma)^{-1}\mathbf{D}_t \tag{C.37}$$

where the constant η is defined as

$$\eta = \frac{\zeta\kappa}{\hat{\rho}} \tag{C.38}$$

C.5 Proof of Corollary 1

Define $R_{i,t}$ as the undiscounted cumulative return of firm i. We have

$$dR_{i,t} = v_i^\top\frac{d\bar{\delta}_t}{V_{i,t}} = \mathbf{e}_i^\top(I - \frac{\beta}{\hat{\rho}}\Gamma)^{-1}\frac{\zeta\kappa}{\hat{\rho}}\frac{d\bar{\delta}_t}{V_{i,t}} = \frac{\zeta\kappa}{\hat{\rho}}\sum_{n=1}^{N}\mathbf{e}_i^\top\left(I - \frac{\beta}{\hat{\rho}}\Gamma\right)^{-1}\mathbf{e}_n\frac{d\delta_{n,t}}{V_{i,t}} \tag{C.39}$$

We are interested in the correlation between $dR_{i,t}, dR_{j,t}$, that is

$$\rho_{ij} = \text{corr}(dR_{i,t}, dR_{j,t}) = \frac{\text{cov}(dR_{i,t}, dR_{j,t})}{\sqrt{\text{var}(dR_{i,t})\text{var}(dR_{j,t})}} \tag{C.40}$$

In calculating covariance, drift terms vanish, and only the diffusion terms contribute. The diffusion term for firm i is $\sigma_{i,\delta}\delta_{i,t}dz_{i,t}$, where $dz_{i,t}$ are i.i.d. stochastic increments with $\text{cov}(dz_{i,t}, dz_{j,t}) = 0$ for $i \neq j$.

Denote $\mathbf{M} = (I - \frac{\beta}{\hat{\rho}}\Gamma)^{-1}$, then $\mathbf{M}$ is a matrix of size $N \times N$. We also denote $\mathbf{M}_i$ as a vector

consisting of i-th row elements of $\mathbf{M}$. That is, $\mathbf{M}_i$ is given by:

$$\mathbf{M}_i = \begin{pmatrix} M_{i,1} \\ M_{i,2} \\ \cdots \\ M_{i,N} \end{pmatrix},$$

where $M_{i,j}$ is the element in the i-th row and j-th column of $\mathbf{M}$. Therefore,

$$\text{corr}(dR_{i,t}, dR_{j,t}) = \frac{\text{cov}\left(\sum_{n=1}^{N} \mathbf{e}_i^\top \mathbf{M} \mathbf{e}_n \frac{d\delta_{n,t}}{V_{i,t}}, \sum_{n=1}^{N} \mathbf{e}_j^\top \mathbf{M} \mathbf{e}_n \frac{d\delta_{n,t}}{V_{j,t}}\right)}{\sqrt{\text{var}\left(\sum_{n=1}^{N} \mathbf{e}_i^\top \mathbf{M} \mathbf{e}_n \frac{d\delta_{n,t}}{V_{i,t}}\right) \text{var}\left(\sum_{n=1}^{N} \mathbf{e}_j^\top \mathbf{M} \mathbf{e}_n \frac{d\delta_{n,t}}{V_{j,t}}\right)}} \tag{C.41}$$

$$= \frac{\text{cov}\left(\sum_{n=1}^{N} \mathbf{e}_i^\top \mathbf{M} \mathbf{e}_n \sigma_{n,\delta} \delta_{n,t} dz_{n,t}, \sum_{n=1}^{N} \mathbf{e}_j^\top \mathbf{M} \mathbf{e}_n \sigma_{n,\delta} \delta_{n,t} dz_{n,t}\right)}{\sqrt{\text{var}\left(\sum_{n=1}^{N} \mathbf{e}_i^\top \mathbf{M} \mathbf{e}_n \sigma_{n,\delta} \delta_{n,t} dz_{n,t}\right) \text{var}\left(\sum_{n=1}^{N} \mathbf{e}_j^\top \mathbf{M} \mathbf{e}_n \sigma_{n,\delta} \delta_{n,t} dz_{n,t}\right)}} \tag{C.42}$$

To further simplify it, we define $\mathbf{dz}$ as the vector of $\sigma_{n,\delta} \delta_{n,t} dz_{n,t}$:

$$\mathbf{dz} = \begin{pmatrix} \sigma_{1,\delta} \delta_{1,t} dz_{1,t} \\ \sigma_{2,\delta} \delta_{2,t} dz_{2,t} \\ \vdots \\ \sigma_{N,\delta} \delta_{N,t} dz_{N,t} \end{pmatrix}.$$

Thus, the summation terms in the numerator of (C.42) become:

$$\sum_{n=1}^{N} \mathbf{e}_i^\top \mathbf{M} \mathbf{e}_n \sigma_{n,\delta} \delta_{n,t} dz_{n,t} = \mathbf{M}_i^\top \mathbf{dz}, \quad \sum_{n=1}^{N} \mathbf{e}_j^\top \mathbf{M} \mathbf{e}_n \sigma_{n,\delta} \delta_{n,t} dz_{n,t} = \mathbf{M}_j^\top \mathbf{dz}.$$

Substitute the vector definitions into the covariance expression:

$$\text{cov}(\mathbf{M}_i^\top \mathbf{dz}, \mathbf{M}_j^\top \mathbf{dz}) = \mathbb{E}[(\mathbf{M}_i^\top \mathbf{dz})(\mathbf{M}_j^\top \mathbf{dz})] - \mathbb{E}[\mathbf{M}_i^\top \mathbf{dz}] \mathbb{E}[\mathbf{M}_j^\top \mathbf{dz}].$$

Since $\mathbf{dz}$ is a random vector with zero mean, we can express the covariance as:

$$\text{cov}(\mathbf{M}_i^\top \mathbf{dz}, \mathbf{M}_j^\top \mathbf{dz}) = \mathbf{M}_i^\top \text{cov}(\mathbf{dz}) \mathbf{M}_j = \mathbf{M}_i^\top \Sigma_z \mathbf{M}_j dt,$$

where $\Sigma_z dt = \text{cov}(\mathbf{dz})$ is the covariance matrix of $\mathbf{dz}$. And because $dz_{i,t}$ is i.i.d., Σ_z is a diagonal

matrix,

$$\Sigma_z = \begin{pmatrix} c_1^2 & 0 & 0 & \cdots & 0 \\ 0 & c_2^2 & 0 & \cdots & 0 \\ 0 & 0 & c_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & c_N^2 \end{pmatrix}.$$

with each diagonal element c_i defined as:

$$c_i^2 = \sigma_{i,\delta}^2 \delta_{i,t}^2, \quad i = 1, 2, \ldots, N.$$

Similarly, the variances terms in the denominator of (C.42) can be simplified as:

$$\text{var}(\mathbf{M}_i^\top \mathbf{dz}) = \mathbf{M}_i^\top \Sigma_z \mathbf{M}_i dt.$$

$$\text{var}(\mathbf{M}_j^\top \mathbf{dz}) = \mathbf{M}_j^\top \Sigma_z \mathbf{M}_j dt.$$

Substitute the covariance and variance expressions into the correlation term:

$$\rho_{ij} = \text{corr}(dR_{i,t}, dR_{j,t}) = \frac{\mathbf{M}_i^\top \Sigma_z \mathbf{M}_j}{\sqrt{\mathbf{M}_i^\top \Sigma_z \mathbf{M}_i \cdot \mathbf{M}_j^\top \Sigma_z \mathbf{M}_j}}.$$

To proceed further we define the following quantity:

$$\|\mathbf{M}_i\| = \sqrt{\mathbf{M}_i^\top \Sigma_z \mathbf{M}_i} = \sqrt{\sum_{n=1}^{N} M_{i,n}^2 c_n^2}.$$

Then the correlation can be expressed as

$$\rho_{ij} = \frac{\mathbf{M}_i^\top \Sigma_z \mathbf{M}_j}{\|\mathbf{M}_i\| \|\mathbf{M}_j\|} \tag{C.43}$$

Next, we characterize the derivative $\frac{\partial \rho_{ij}}{\partial \Gamma_{ij}}$. We will use chain rule to calculate its derivative for its numerator and its denominator separately.

First note that, from matrix calculus, for $\mathbf{M} = (I - \frac{\beta}{\hat{\rho}}\Gamma)^{-1}$:

$$\frac{\partial \mathbf{M}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} \mathbf{M} E_{ij} \mathbf{M} \tag{C.44}$$

where E_{ij} is the elementary matrix with 1 at (i, j) and 0 otherwise. This implies, for any

$k, l \in \{1, 2, ..., N\}$:

$$\frac{\partial M_{k,l}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} (\mathbf{M} E_{ij} \mathbf{M})_{k,l} = \frac{\beta}{\hat{\rho}} M_{k,i} M_{j,l}. \tag{C.45}$$

Numerator Now let us first focus on the numerator:

$$S = \mathbf{M}_i^\top \Sigma_z \mathbf{M}_j = \sum_n c_n^2 M_{i,n} M_{j,n}.$$

Expanding its derivative with respect to Γ_{ij} using the product rule:

$$\frac{\partial S}{\partial \Gamma_{i,j}} = \sum_n \left(\frac{\partial M_{i,n}}{\partial \Gamma_{i,j}} M_{j,n} c_n^2 + c_n^2 M_{i,n} \frac{\partial M_{j,n}}{\partial \Gamma_{i,j}} \right).$$

We use (C.45) to calculate the derivatives:

$$\frac{\partial M_{i,n}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} M_{i,i} M_{j,n}, \tag{C.46}$$

$$\frac{\partial M_{j,n}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} M_{j,i} M_{j,n}. \tag{C.47}$$

Substituting these derivatives gives:

$$\frac{\partial S}{\partial \Gamma_{i,j}} = \sum_n \frac{\beta}{\hat{\rho}} c_n^2 \left(M_{i,i} M_{j,n}^2 + M_{j,i} M_{i,n} M_{j,n} \right).$$

Using the definition of $\|\mathbf{M}_j\|^2$ and S:

$$\frac{\partial S}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} \left(M_{i,i} \|\mathbf{M}_j\|^2 + M_{j,i} S \right).$$

Denominator We next calculate

$$\frac{\partial}{\partial \Gamma_{i,j}} (\|\mathbf{M}_i\| \|\mathbf{M}_j\|) = \|\mathbf{M}_j\| \frac{\partial \|\mathbf{M}_i\|}{\partial \Gamma_{i,j}} + \|\mathbf{M}_i\| \frac{\partial \|\mathbf{M}_j\|}{\partial \Gamma_{i,j}}.$$

Since $\|\mathbf{M}_i\| = \sqrt{\sum_n c_n^2 M_{i,n}^2}$, we have:

$$\frac{\partial \|\mathbf{M}_i\|}{\partial \Gamma_{i,j}} = \frac{1}{2\|\mathbf{M}_i\|} \frac{\partial}{\partial \Gamma_{i,j}} \left(\sum_n c_n^2 M_{i,n}^2 \right).$$

The derivative of the summation term is:

$$\frac{\partial}{\partial \Gamma_{i,j}} \left(\sum_n c_n^2 M_{i,n}^2 \right) = 2 \sum_n c_n^2 M_{i,n} \frac{\partial M_{i,n}}{\partial \Gamma_{i,j}}.$$

Substitute $\frac{\partial M_{i,n}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} M_{i,i} M_{j,n}$:

$$\frac{\partial \|\mathbf{M}_i\|}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} \frac{M_{i,i}}{\|\mathbf{M}_i\|} \sum_n c_n^2 M_{i,n} M_{j,n} = \frac{\beta}{\rho} \frac{M_{i,i} S}{\|\mathbf{M}_i\|}.$$

Similarly:

$$\frac{\partial \|\mathbf{M}_j\|}{\partial \Gamma_{i,j}} = \frac{1}{2\|\mathbf{M}_j\|} \frac{\partial}{\partial \Gamma_{i,j}} \left(\sum_n c_n^2 M_{j,n}^2 \right).$$

The derivative of the summation term is:

$$\frac{\partial}{\partial \Gamma_{i,j}} \left(\sum_n c_n^2 M_{j,n}^2 \right) = 2 \sum_n c_n^2 M_{j,n} \frac{\partial M_{j,n}}{\partial \Gamma_{i,j}}.$$

Substitute $\frac{\partial M_{j,n}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} M_{j,i} M_{j,n}$:

$$\frac{\partial \|\mathbf{M}_j\|}{\partial \Gamma_{i,j}} = \frac{\frac{\beta}{\rho} M_{j,i}}{\|\mathbf{M}_j\|} \sum_n c_n^2 M_{j,n} M_{j,n} = \frac{\beta}{\hat{\rho}} M_{j,i} \|\mathbf{M}_j\|.$$

Therefore

$$\frac{\partial}{\partial \Gamma_{i,j}} (\|\mathbf{M}_i\| \|\mathbf{M}_j\|) = \frac{\beta}{\hat{\rho}} \left(\frac{M_{i,i} S \|\mathbf{M}_j\|}{\|\mathbf{M}_i\|} + M_{j,i} \|\mathbf{M}_i\| \|\mathbf{M}_j\| \right).$$

Substituting these results:

$$\frac{\partial \rho_{ij}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} \frac{\left(M_{i,i} \|\mathbf{M}_j\|^2 + M_{j,i} S \right) \|\mathbf{M}_i\| \|\mathbf{M}_j\| - S \cdot \left(\frac{M_{i,i} S \|\mathbf{M}_j\|}{\|\mathbf{M}_i\|} + M_{j,i} \|\mathbf{M}_i\| \|\mathbf{M}_j\| \right)}{\|\mathbf{M}_i\|^2 \|\mathbf{M}_j\|^2}.$$

Simplifying the numerator and factoring out $M_{i,i} \|\mathbf{M}_j\| / \|\mathbf{M}_i\|$, the numerator becomes:

$$\frac{\beta}{\hat{\rho}} \frac{M_{i,i} \|\mathbf{M}_j\|}{\|\mathbf{M}_i\|} \left(\|\mathbf{M}_i\|^2 \|\mathbf{M}_j\|^2 - S^2 \right).$$

Thus, the final expression for the derivative is:

$$\frac{\partial \rho_{ij}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} \frac{M_{i,i} \left(\|\mathbf{M}_j\|^2 \|\mathbf{M}_i\|^2 - S^2 \right)}{\|\mathbf{M}_i\|^3 \|\mathbf{M}_j\|}.$$

Since S is a inner product, we can apply the Cauchy-Schwarz inequality:

$$\left(\sum_{n=1}^{N} c_n^2 M_{i,n} M_{j,n} \right)^2 \leq \left(\sum_{n=1}^{N} c_n^2 M_{i,n}^2 \right) \left(\sum_{n=1}^{N} c_n^2 M_{j,n}^2 \right).$$

which simplifies to:

$$S^2 \leq \|\mathbf{M}_i\|^2 \|\mathbf{M}_j\|^2.$$

Thus, we obtain:

$$\|\mathbf{M}_i\|^2 \|\mathbf{M}_j\|^2 - S^2 \geq 0.$$

which implies:

$$\frac{\partial \rho_{ij}}{\partial \Gamma_{i,j}} > 0.$$

The inequality holds strictly because $\mathbf{M}_i$ and $\mathbf{M}_j$ are not collinear. Moreover, we see that when $\beta = 0$, the network effect disappears entirely. As β increases, it amplifies the overall network influence—specifically, a higher β strengthens the network's impact, raising the sensitivity of the correlation with respect to $\Gamma_{i,j}$. That is, $\frac{\partial^2 \rho_{ij}}{\partial \Gamma_{i,j} \partial \beta} > 0$.

C.6 Proof of Proposition 3

For firm i, substituting the functional form of its valuation (C.21) into the HJB equation, then taking the FOC with respect to $x_{i,t}$, we get

$$x_{i,t} = \frac{\phi_0}{\phi_1} D_{i,t} - \frac{\zeta}{\theta v_{i,i}} \tag{C.48}$$

Where $v_{i,i}$ is the i-th element of vector v_i. $\phi_0 > 0, \phi_1 > 0$, so $x_{i,t}$ is increasing with $D_{i,t}$. And v_i is solved from (C.34). Recall that the law of motion of data capital is given by

$$d\delta_{i,t} = \left(\theta x_{j,t} + \alpha \theta D_{j,t} + \mu_{i,t} \delta_{i,t} \right) dt + \sigma_{i,\delta} \delta_{i,t} dz_{i,t} \tag{C.49}$$

Consequently, taking the difference of firm product design choice $x_{i,t}$, we obtain

$$dx_{i,t} = \frac{\phi_0}{\phi_1} \left[\sum_{j=1}^{N} \gamma_{ij} \left((\theta x_{j,t} + \alpha \theta D_{j,t} + \mu_{j,\delta} \delta_{j,t}) dt + \sigma_{j,\delta} \delta_{j,t} dz_{j,\delta,t} \right) \right] \tag{C.50}$$

C.7 Proof of Proposition 4

Recall that in C.4 we show the firm's valuation takes the following form

$$V_{i,t} = v_{i,0} + v_i^\top \bar{\delta}_t \tag{C.51}$$

Therefore, the firm i's sensitivity to other firms data is given by

$$(\frac{\partial V_{i,t}}{\partial \bar{\delta}})^\top = v_i \tag{C.52}$$

The j-th entry of this vector is given by

$$v_{i,j} = \mathbf{e}_i^\top \left(\mathbf{I} + \frac{\beta}{\hat{\rho}}\Gamma + (\frac{\beta}{\hat{\rho}})^2\Gamma^2 + (\frac{\beta}{\hat{\rho}})^3\Gamma^3 + ... \right) \mathbf{e}_j \tag{C.53}$$

Since all the entry of Γ is positive, and $\beta = \theta(\frac{\phi_0}{\phi_1} + \alpha)$ is increasing in θ, all the entry of v_i are increasing in θ.

C.8 Proof of Proposition 5

Note that the valuation of firm i is given by

$$V_{i,t} = v_{i,0} + \eta\delta_{i,t} + \eta\frac{\beta}{\hat{\rho}}\mathbf{e}_i^\top \sum_{k=0}^{\infty} \left(\frac{\beta}{\hat{\rho}}\Gamma\right)^k (\mathbf{D}_t) \tag{C.54}$$

We can rewrite it as

$$V_{i,t} = v_{i,0} + \eta\delta_{i,t} + \eta\frac{\beta}{\hat{\rho}}\mathbf{e}_i^\top \left(\mathbf{I} + \frac{\beta}{\hat{\rho}}\Gamma + (\frac{\beta}{\hat{\rho}})^2\Gamma^2 + ... \right) \Gamma\bar{\delta}_t \tag{C.55}$$

$$= v_{i,0} + \eta\mathbf{e}_{i,t}^\top\mathbf{I}\bar{\delta}_t + \eta\frac{\beta}{\hat{\rho}}\mathbf{e}_i^\top \left(\Gamma + \frac{\beta}{\hat{\rho}}\Gamma^2 + (\frac{\beta}{\hat{\rho}})^2\Gamma^3 + ... \right) \bar{\delta}_t \tag{C.56}$$

$$= v_{i,0} + \eta\mathbf{e}_i^\top \left(\mathbf{I} + \frac{\beta}{\hat{\rho}}\Gamma + (\frac{\beta}{\hat{\rho}})^2\Gamma^2 + (\frac{\beta}{\hat{\rho}})^3\Gamma^3 + ... \right) \bar{\delta}_t \tag{C.57}$$

$$= v_{i,0} + \zeta\kappa\mathbf{e}_i^\top(\hat{\rho}\mathbf{I} - \beta\Gamma)^{-1}\bar{\delta}_t \tag{C.58}$$