Data as a Networked Asset*

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

August 16, 2025

Data is non-rival: a firm's customer data informs other firms about their customers. We uncover a network of inter-firm data conduits embedded in mobile applications. Data sharing induces comovement in firms' operational, financial, and stock-market performances, propagates shocks (e.g., cyberattacks), and induces herding in product design. Apple's privacy policy—a shock to inter-firm data flows—weakened these patterns. We develop a dynamic network model, where firms' performance and growth are interconnected through a data-sharing network. A network-augmented Gordon growth formula emerges for valuing data-generated cash flows. Our valuation metrics incorporate high-order and long-term spillovers and reveal systemically important firms.

Keywords: Data, data sharing, network, systemic importance, customer capital, intangible, privacy, cyberattack

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

*We thank Philip Bond, Cecilia Bustamante, Daniel Chen, Tony Cookson, Nicolas Crouzet, Janice Eberly, Maryam Farboodi, Naveen Gondhi, Jerry Hoberg, Allen Hu, Kai Li, Yi Li, Laura Xiaolei Liu, Dimitris Papanikolaou, Joel Peress, Adriano Rampini, Amit Seru, Zheng (Michael) Song, Eduard Talamàs, Luke Taylor, Di Tian, Laura Veldkamp, Xiaoyun Yu, Shaojun Zhang and conference and seminar participants at the Barcelona Summer Forum, Boston College, CICF, FIRS, Future of Financial Information Conference, Maryland Junior Finance Conference, Peking University, RCFS Winter Finance Conference, SAIF Annual Research Conference, SFS Cavalcade, the 11th Biennial Conference on Transition and Economic Development, UNC-Duke Corporate Finance Conference, University of Chinese Academy of Sciences, and University of Macau 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 and the Social Sciences and Humanities Research Council of Canada (Grant Number: 435-2025-0700). 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 used by multiple firms simultaneously (Jones and Tonetti, 2020). Data also exhibits externalities: information collected by one firm on its customers can benefit other firms by revealing consumer preferences, underlying economic forces, and serving as training data for prediction models (Choi, Jeon, and Kim, 2019; Ichihashi, 2021; Acemoglu, Makhdoumi, Malekian, and Ozdaglar, 2022). Data non-rivalry, combined with externality, expands 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 and network structure of data sharing? What are the economic implications, and in particular, how data sharing affects firms' decision-making and propagates shocks across firms? What is the impact and unintended consequences of data regulations? 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 inter-firm data flows 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 Compustat firms. Firms may operate one or multiple mobile apps. These apps collect data and transmit it to Software Development Kits (SDKs) that specialize in data aggregation and analytics, which we refer to as "data-sharing SDKs". A crucial function of these 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 these 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 the usage of data-sharing SDKs. Specifically, two firms are considered connected when a common set of such 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 data-sharing conduits between firms and to map out the entire network structure.

The signals that connected firms receive from data-sharing SDKs contain data from one another. When one firm gains 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 customers and improve profitability. Our empirical analysis confirms this dynamic. We find that data sharing induces 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 factors or other common exposures, such as product similarity, supply-chain linkages, 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). ATT allows app users to block identity tracking, thus making it difficult for the SDKs to merge data from different firms.

By examining the impact of ATT, we not only shed light on the consequences of privacy initiatives from the private sector but also validate that firms' overlap in data-sharing SDKs is fundamentally about linkages in data. Importantly, ATT does not influence the contribution of other forms of firms' overlap (e.g., product similarity) to their performance comovements.

The comovement in firms' performances emerges ultimately from the propagation of shocks from one firm to another. Using cyberattacks as our empirical setting, we provide direct evidence of how data sharing facilitates shock propagation. Intuitively, cyberattacks reduce firms' data stock and access, impair their ability to collect data, and more broadly, diminish operational efficiency that underpins customer engagement and data acquisition. We find that the focal firm's data-connected peers experience significantly greater deterioration in operational and stock-market performances than those that are not data-connected. These findings underscore the importance of examining firms' interconnectedness in the data economy for tracing the ripple effects of shocks.

Given the network structure of data sharing, a natural question arises: which firms are systemically important in the data economy? Data from systemically important firms exerts a disproportionate impact on the aggregate data ecosystem due to these firms' critical positions in the network. Answering this question requires accounting not only for direct spillover effects—which have been the focus of our empirical analysis thus far—but also for higher-order spillovers. For example, firm A's data may be shared via SDKs with firm B, which in turn shares data with firm C. Furthermore, data spillovers have persistent effects across multiple time horizons: data influences firms' ability to engage with customers, and customer engagement, in turn, drives further data accumulation, resulting in self-perpetuating growth that is interconnected across firms.

To assess firms' systemic importance within this networked data economy, we develop a dynamic network model where data plays the role of productive capital. The model takes as input the empirical network of data-sharing conduits, and simulations of the calibrated model replicate the data-induced comovement in firm performances and the ATT effects. In the model, each firm

seeks to forecast customer preferences, with forecasting precision depending on both its own data and data from connected peers.[1] More accurate forecasts enhance customer engagement, which in turn boosts customer capital, product demand, and cash flow.[2] Customer engagement also creates new data, improving forecasting precision for both the firm itself and its connected peers.

Building on the dynamic model, we derive a network-augmented Gordon growth formula for firm valuation—that is, the present value of a firm's cash flows, which accounts for the impact of both direct and indirect data spillovers from other firms and the self-perpetuating data growth over all time horizons. A firm's valuation (i.e., value function) depends on its own data stock as well as the data stocks of its connected peers (the state variables of the economy). Firms' valuation systematically captures how data propagates through the network to influence long-run cash-flow dynamics. In the absence of data sharing, our formula collapses to the standard Gordon growth model. Our model provides a tractable approach for valuing data as a networked productive asset.

Importantly, the valuation formula also provides a framework for quantifying firms' systemic importance. We decompose the aggregate valuation of all firms into individual firms' contributions, recognizing that each firm's value depends not only on its own data but also on data from others. As a result, a firm's contribution to aggregate valuation can differ substantially from its own standalone valuation. For example, prior to the introduction of ATT, Meta (formerly Facebook) had a contribution-to-valuation ratio of 4.2, meaning that removing Meta from the data economy would reduce aggregate cash flows by an amount equivalent to 4.2 times its own valuation. After ATT, this ratio declined to 1.2. These findings underscore that under data sharing, certain firms become systemically important in ways not fully reflected by their market value or size alone.

Finally, our model features a trade-off that is distinctive to firms in the data economy. When designing products, a firm balances monetization and customer engagement. For instance, a software company offers services for free to boost customer engagement, but doing so reduces current cash flows. We characterize an intertemporal trade-off: prioritizing customer engagement and data collection contributes to future knowledge on customers and future profitability but compromises current profits. Data from connected peers informs the firm about its customers and thereby alleviates the tension. In the example, being more informed about its customers allows the software company to target those who are willing to pay for the premium features and extract more profits.

Intriguing product-design dynamics emerge in our model. When one firm prioritizes customer engagement, it generates more data that informs its connected peers about their customers,

[1] In Buera and Oberfield (2020), there is randomness in a firm's adoption of insights from other industries. In our model, a firm's information on its customers depends on a weighted sum of other firms' data, with the weight corresponding to the empirically measured data-sharing conduits. While the amount of data transmitted via the conduits varies over time and carries randomness (shocks), the data-sharing conduits are deterministic.

[2] Customer capital accumulates, for example, as data improves firm-consumer match (Gourio and Rudanko, 2014b). Hsieh and Rossi-Hansberg (2023) point out that intangible capital lowers the cost of entering new markets and acquiring customers. Data, as a particular form of intangibles, serves a similar purpose as it facilitates customer acquisition.

thus allowing these peer firms to prioritize customer engagement (and generate more data as well) without significantly sacrificing current profits. Conversely, when a firm downplays customer engagement (prioritizes monetization), its reduced data collection and diminished data spillover make it harder for the connected peers to balance profitability and customer engagement. For any given level of profitability, these connected peer firms, now being less informed about their customers, have to sacrifice more customer engagement in their product design 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 and supply-chain linkages. Furthermore, herding in product design among data-connected firms is weakened by the introduction of ATT, indicating that this empirical pattern is indeed driven by data sharing.[3]

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.[4] 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.

Literature. To the best of our knowledge, our paper is the first to systematically measure the data-sharing network and explore its economic implications. Data is non-rival. It can be used by multiple firms at the same time (e.g., Jones and Tonetti, 2020). The data conduits based on shared SDKs in firms' mobile applications have become the primary channels for inter-firm data sharing. We embed this empirically measured network of data-sharing conduits into an otherwise canonical model of data-driven firm growth (e.g., Farboodi and Veldkamp, 2021).

Data is an intangible asset. The network of data-sharing conduits reflects the scope of data usage. Our paper contributes to the literature on the scope of intangible asset usage.[5] While

[3]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.

[4]For example, investment by one firm may increase the cost of investment inputs and financing or intensify product-market competition that other firms face (e.g., Asriyan, Laeven, Martin, Van der Ghote, and Vanasco, 2024).

[5]More broadly, our paper contributes to the literature on the non-rival nature of intangible capital (e.g., McGrattan

the existing papers focus on the scope of intangible usage within firms (e.g., Argente, Moreira, Oberfield, and Venkateswaran, 2021; Crouzet, Eberly, Eisfeldt, and Papanikolaou, 2024), our paper characterizes the usage of data across firms and explores the implications on firm growth, a central topic in the literature, and systemic risk, which is a new question that we connect to the scope of intangible usage: data sharing facilitates growth but also propagates shocks by synchronizing firms' behavior, thereby amplifying aggregate fluctuations.[6] When it comes to data usage, our paper emphasizes data access rather than data ownership, following Varian (2019). In our setting, data is generated as a by-product of firms' operations (e.g., Bergemann, Bonatti, and Gan, 2022; Farboodi and Veldkamp, 2021).[7] Firms take as given the network of data-sharing conduits. Data is shared automatically, spurring interconnected firm growth and shock propagation. Beyond the scope of our paper, data ownership is often discussed in association with market-based approaches to data sale (e.g., Acemoglu, Makhdoumi, Malekian, and Ozdaglar, 2022; Liu, Ma, and Veldkamp, 2025) and related to compensating data contributors, particularly in response to privacy concerns.

Our paper contributes to the literature on intangible asset valuation.[8] Empirically, we show that firms' stock-return comovements can be attributed to data-sharing linkages. Theoretically, our network-augmented Gordon growth formula is a tractable framework for valuing data-generated cash flows. Valuing data as a productive asset has become increasingly important (Veldkamp, 2023). Traditional cost-based methods face clear limitations as the cost of data acquisition is not well defined when data is a by-product of firms' operations. Prior studies measure the value of data indirectly by focusing on complementary labor inputs (e.g., Abis and Veldkamp, 2023; Corhay, Hu, Li, Tong, and Tsou, 2024) or the outcome of data usage (e.g., Eeckhout and Veldkamp, 2022; Farboodi, Singal, Veldkamp, and Venkateswaran, 2024; Dong, Hu, Li, and Liu, 2025).

In our model, firms' decision-making and valuation are interconnected through a network adjacency matrix of data conduits, and the equilibrium conditions have a spatial structure (Comin, Dmitriev, and Rossi-Hansberg, 2012; de Paula, 2017; Redding and Rossi-Hansberg, 2017). Following Diebold and Yılmaz (2014), Ballester, Calvo-Armengol, and Zenou (2006), and Denbee, Julliard, Li, and Yuan (2021), we decompose aggregate valuation of data-generated cash flows into individual firms' contributions and develop the first metric of firms' systemic importance in the data economy.[9] This approach of modeling systemically important agents has been adopted in

and Prescott, 2009, 2010; Cong, Xie, and Zhang, 2021; Crouzet, Eberly, Eisfeldt, and Papanikolaou, 2022).

[6]Our mechanism is different from Veldkamp and Wolfers (2007) who emphasize costly information acquisition.

[7]The notion that data is a by-product of economic activity was well established in the information economics literature (e.g., Veldkamp, 2005; Ordoñez, 2013; Fajgelbaum, Schaal, and Taschereau-Dumouchel, 2017).

[8]There is a growing literature on measuring and valuing intangible capital (e.g., Eisfeldt and Papanikolaou, 2013; Gourio and Rudanko, 2014b; Kogan, Papanikolaou, Seru, and Stoffman, 2017; Peters and Taylor, 2017; Kelly, Papanikolaou, Seru, and Taddy, 2021; Bhandari and McGrattan, 2021; Dou, Ji, Reibstein, and Wu, 2021; **?**). The cost-based approach has been a primary method in the literature.

[9]Our paper adds to the systemic risk literature (Billio, Getmansky, Lo, and Pelizzon, 2012; Acharya, Pedersen, Philippon, and Richardson, 2016; Adrian and Brunnermeier, 2016; Benoit, Colliard, Hurlin, and Pérignon, 2016; Bai,

studies on social connections (e.g., Graham, 2008; Calvó-Armengol, Patacchini, and Zenou, 2009; Fogli and Veldkamp, 2021) and financial markets (e.g., Ozdagli and Weber, 2017; Herskovic, 2018; Eisfeldt, Herskovic, Rajan, and Siriwardane, 2022; Eisfeldt, Herskovic, and Liu, 2023).

Our findings on the economic implications of data sharing, the ATT impact, and cyberattack contribute to several strands of the empirical literature. Data regulations restrict the scope of data usage, thereby impeding firm growth. The associated negative impact on firms' valuations is in line with the direct evidence on how ATT affects firms' stock prices (Bian, Ma, and Tang, 2021). Our paper contributes by uncovering an unintended benefit of data regulations: they reduce systemic risk by weakening the comovement of firms' performance and herding in firms' product design.

Prior studies have examined how data regulations affect firms, including outcomes such as web traffic (Goldberg, Johnson, and Shriver, 2024), revenues (Aridor, Che, and Salz, 2023), innovation and venture investment (Bessen, Impink, Reichensperger, and Seamans, 2020; Janssen, Kesler, Kummer, and Waldfogel, 2022; Jia, Jin, and Wagman, 2021), SDK usage in Android mobile apps (Jin, Liu, and Wagman, 2024), data reliance (Demirer, Jiménez Hernández, Li, and Peng, 2024), and firms' technology choice (Peukert, Bechtold, Batikas, and Kretschmer, 2022).[10]

Few empirical studies examine spillover effects in the data economy. Aridor, Che, and Salz (2023) document consumer-side spillovers: consumers' privacy decisions enable firms to infer other consumers' types. Using ATT as a shock to the ease of merging data across firms, our paper complements this literature 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 provides new evidence on the economic implications of cyberattacks. Crosignani, Macchiavelli, and Silva (2023) document the propagation of cyberattacks through supply chains. Akey, Lewellen, Liskovich, and Schiller (2023) examine the impact of cyberattacks on firm value. Our paper adds to this literature by documenting how the negative impact of cyberattacks is propagated across firms through the data-sharing network.

In summary, our empirical analysis reveals that firms are not only interconnected through the conventional economic relationships—such as supply chains and production networks (e.g., Cohen and Frazzini, 2008; Menzly and Ozbas, 2010; Barrot and Sauvagnat, 2016; Auer, Levchenko, and Sauré, 2019; Boehm, Flaaen, and Pandalai-Nayar, 2019; Carvalho, Nirei, Saito, and Tahbaz-Salehi, 2021; Huo, Levchenko, and Pandalai-Nayar, 2025), product-market overlap (e.g., Hoberg and Phillips, 2010, 2016, 2018), locations (e.g., Parsons, Sabbatucci, and Titman, 2020; Comin,

Krishnamurthy, and Weymuller, 2018; Duarte and Eisenbach, 2021; Greenwood, Landier, and Thesmar, 2015).

[10]Specifically about ATT and ban on the use of third-party cookies, the literature has examined outcomes including firms' product monetization decisions (Kesler, 2023; Aridor, Che, Hollenbeck, Kaiser, and McCarthy, 2024), advertising effectiveness (Alcobendas, Kobayashi, Shi, and Shum, 2023; Aridor et al., 2024; Wernerfelt, Tuchman, Shapiro, and Moakler, 2024), app market concentration (Li and Tsai, 2022), financial forecasting (Abis, Tang, and Bian, 2025), and fraud (Bian, Pagel, Tang, and Raval, 2023). About GDPR, Johnson (2022) provides a review of the literature.

Dmitriev, and Rossi-Hansberg, 2012), technological proximity (e.g., Bloom, Schankerman, and Van Reenen, 2013; Liu and Ma, 2021), or shared analyst coverage (e.g., Ali and Hirshleifer, 2020)—but also by the data-sharing conduits enabled by the SDKs embedded in firms' mobile applications. Our contribution lies in documenting how this new type of economic linkage impacts firms' performance comovement, corporate valuation, and shock propagation across firms.

2 Data Sharing Network: Measurement and Evidence

2.1 The empirical setting

Background: data economy and mobile applications. Digital economy is 10.0% of U.S. GDP and has an annualized growth rate of 7.1% from 2017 to 2022.[11] A key driver of the digital economy's rise is the growing use of mobile devices—a trend that accelerated during the COVID-19 pandemic when demand for digital services across work, entertainment, and communication skyrocketed. In 2022, the average U.S. adult spent over 4.5 hours per day on mobile devices.[12]

Data is an essential productive asset in the digital economy. Companies rely on data collected from mobile devices to understand consumer preferences, customize product offerings, and guide innovation choices. Bian et al. (2021) find that over 60% of mobile applications (apps) tracks users across websites, apps, and offline stores. Binns, Lyngs, Van Kleek, Zhao, Libert, and Shadbolt (2018) report that nearly 90% of Android apps collect user data and enable data-sharing with Google. Companies leveraging consumer data for targeted advertising—such as Google and Meta—generated approximately $780 billion in revenue in 2023.[13]

A key advantage of mobile device data is its high connectivity, enabled by consistent and universal identifiers such as the Identifier for Advertisers (IDFA). This device-level identifier is assigned by Apple and allows app developers, advertisers, analytics platforms, and ad networks to track user behavior across iOS apps, providing a consistent framework for linking activity and building user profiles from diverse data sources.[14] 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. In addition, users can delete or block cookies in their browser settings. Overall, IDFA offers greater consistency, making mobile data a valuable asset for firms.

Firms in the data economy. Motivated by the growing importance of app-based data collection and utilization, we analyze firms with at least one mobile app. Our empirical exercises require in-

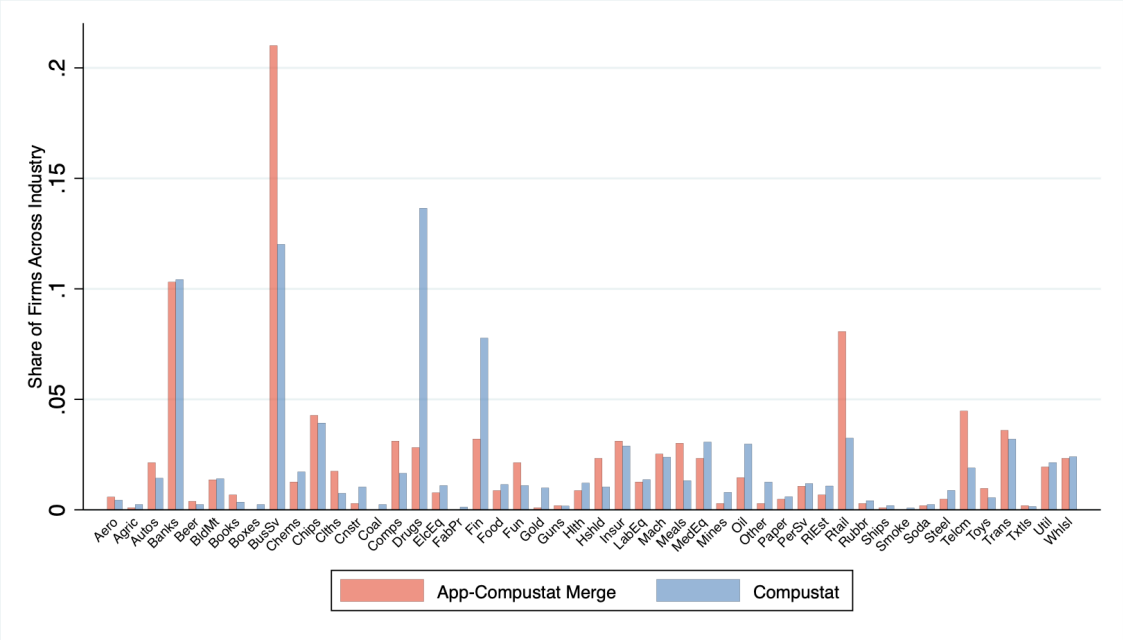
[11] Source: https://www.bea.gov/data/special-topics/digital-economy

[12] Source: https://www.emarketer.com/content/us-time-spent-with-connected-devices-2022

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

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

FIGURE 1: **Firms in Our Sample vs. Compustat Universe: Industry Distribution**



NOTE.—Figure 1 compares the industry distribution of firms in the data economy ("App-Compustat Merge") with those in the broader Compustat universe ("Compustat"). It shows the distribution over the Fama-French 48 industries.

formation on firms' mobile apps and standard accounting information from Compustat. A total of 1,031 firms meet the criteria. They cover over 60% of the total assets of Compustat firms. The distribution of firms across industries is comparable between our sample and the broader Compustat sample (see Figure 1). Certain industries, such as Business Services, Retail, and Telecommunications, are over-represented in our sample, while several others, such as Finance and Oil, are under-represented. In the Internet Appendix, we provide additional information on firms in our sample. Figure A.1 shows the share of firms that our sample covers relative to the broader Compustat universe, as measured by both firm count and total assets. In 24 out of the 48 Fama-French industries, our sample firms account for more than 50% of total assets. In Figure A.2, we examine the distribution of firm size (logarithm of total assets) in Panel A and asset turnover (sales-to-asset ratio) in Panel B. Firms in our sample are larger and more efficient in generating revenues.

SDK and data sharing. We infer the data-sharing network from firms' usage of Software Development Kits (SDKs). SDKs are pre-built app components ("building blocks") that firms integrate into their mobile apps. SDKs support various functionalities, such as messaging and payment. A subset of SDKs facilitates data sharing. They collect user data from host apps. By using unique identifiers such as IDFA, these SDKs link user activity across multiple apps, enabling the construction of comprehensive consumer profiles. Firms integrate such SDKs in their apps to access these customer profiles, which provide valuable insights for marketing, product development, and other

strategic decisions.[15] Therefore, an SDK is a data processing center: it takes as input the data from all firms that install it on their mobile apps and delivers customer profiles as output.

Two firms that share data with the same SDKs are thus connected as the customer profiles they obtain from the SDKs contain data from both firms. Even though firms may not share the same product categories, information on one firm's customers can still be valuable for revealing customer demand of another firm through information on customers' spending patterns and income.

We use Amazon and General Motors (GM) to illustrate data linkages. According to our measure introduced below in Section 2.2, GM exhibits one of the strongest data linkages with Amazon among all firms, because both install a similar set of data-sharing SDKs on their apps—that is, they share the "data conduits." GM collects extensive data on consumer behavior through vehicle purchases, financing, and usage of its IoT (internet of things) software on the vehicles.[16] By sharing data with the SDKs, GM enables the SDKs to form more informative consumer profiles that contribute to Amazon's advertisement targeting. 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. The improved precision in targeting enhances Amazon's ability to convert high-value customers while improving its inventory planning for automotive-related products, boosting overall sales and profitability.

Similarly, data collected by Amazon flows through the SDKs and gets merged into consumer profiles that GM leverages to predict consumer preferences and market trends. As disclosed in the privacy labels of Amazon's iOS mobile app, Amazon collects user data that facilitates third-party advertising. Data on consumer purchasing trends, such as demand for EV-related products among specific demographic groups, helps GM tailor its marketing strategies and product offerings.

The ATT shock. Our empirical analysis leverages a policy shock introduced by Apple's App Tracking Transparency (ATT), implemented on April 26, 2021, which requires app developers to obtain explicit user consent before tracking user activity. It negatively affected firms' data collection and sharing. Specifically, the policy requires apps to display a prompt asking for permission to track user activity. Opt-in rates were low, ranging from 4% to 18% in the first 12 months following the rollout of ATT in the U.S.[17] As a result, ATT substantially limited data collection and the cross-firm linkage of user information based on iOS users' universal identifiers.

[15] SDK providers often generate indirect revenue through advertising. By supplying customer profiles to companies that use their data analytics SDKs, they allow these firms to engage in more effective targeted advertising, and a portion of the resulting advertising expenditures, paid to ad publishers, ultimately flows back to the SDK provider.

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

[17] See more details at `https://www.flurry.com/blog/att-opt-in-rate-monthly-updates/`. Kraft, Skiera, and Koschella (2023) document a 17% opt-in rate as of March 2022.

2.2 Data sources and variable construction

The main data source. We begin by introducing our primary data source on firms' app design and usage. We use two products from Apptopia, a leading provider of mobile app intelligence that has systematically tracked the universe of mobile applications since 2014. First, we obtain information about app characteristics (e.g., category, age) and performance (e.g., downloads, active users, sessions length). Second, we use Apptopia's SDK Intelligence, which provides data on the installation and removal dates of SDKs for each app. More specifically, Apptopia tracks the history of apps' SDK installations and removals in mobile apps by analyzing the publicly available app installation package. When a new version of an app is released, Apptopia reanalyzes the update package to maintain an accurate and current record of the app's SDK profile.

Apptopia classifies SDKs into several categories. We identify the data-sharing SDKs as those that are related to advertising network, marketing campaigns, data analytics, and other activities that analyze and improve customers' willingness to pay. We verify the accuracy of SDKs' functionality by reviewing the documentation files in their GitHub repositories. We focus on the 50 most popular data-sharing SDKs in iOS apps, ranked by worldwide net installations as of April 26, 2021 when ATT was introduced. The most popular SDK accumulated 262,209 net installations (i.e., installations minus uninstallations) since 2014, while the 50$^{\text{th}}$ most popular SDK accumulated only 5,096 net installations in the same period, indicating a high level of concentration. In the Internet Appendix, we conduct robustness analysis by considering the top 20 SDKs.

The remaining data sources are standard. We use Compustat and CRSP for financial information, and draw on the Text-based Network Industry Classifications (TNIC), FactSet Revere, USPTO, and I/B/E/S to construct various measures of firm linkages introduced below. Details of these data sources are provided in the Internet Appendix B.1.

Data connectedness. To characterize the data-sharing network, we develop a pairwise measure for data connectedness. This methodology is akin to the approach taken by Hoberg and Phillips (2010, 2016), which develop a product overlap measure to study competition 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 between pairs of firms based on their usage of data-related SDKs in their iOS apps.[18]

Specifically, we denote app a's (from firm i) data conduit via SDK k at time (quarter) t as: $s_{iakt} = m_{iat} \times d_{iakt}$, where m_{iat} is number of monthly active users (MAU) 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. The installation indicator is scaled by the size of app user base to capture how relevant this data conduit is.

[18]The SDKs used in firms' Android and iOS apps are often different. Since ATT only affects data sharing in the iOS environment, we focus exclusively on data connectedness based on overlaps in SDK usage across iOS apps.

Aggregating across all apps owned by firm i at time t, we define firm i's data conduit via 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 all of firm i's data conduits at time t into a $K \times 1$ vector: $\mathbf{S_{it}} = \left[S_{i1t}, S_{i2t}, ..., S_{iKt}\right]'$, where K is the total number of SDKs. K is 50 in the baseline version of our measure. The cosines-similarity between firm i's and j's data conduits 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. Data connectedness is similar within industry and across industries (see Figure 4). Our measure of data connectedness reveals a new type of firm linkages, distinct from traditional ones within an industry or along the supply chain that we control in our analysis. Appendix C visualizes the entire network structure of data connectedness (see Figure 3).

Performance comovement. We measure the performance comovement between two firms by calculating the correlation across various metrics, including the logarithm of downloads and average daily active users (DAU), earnings growth, and asset turnover (sales/assets), computed at a quarterly frequency.[19] For each performance metric, we compute one correlation for quarters before the implementation of ATT in 2021Q2 and another for those after the implementation. Our full-sample period is 2014Q3–2023Q2. We also consider comovement in firms' stock returns. For each firm pair, comovement is measured as the correlation of monthly returns in a rolling 12-month window from 2014 to 2023.[20] We consider three types of returns: raw returns, abnormal returns based on CAPM, and DGTW-adjusted returns (Daniel, Grinblatt, Titman, and Wermers, 1997).

Other firm linkages. Our focus is on the impact of data connectedness on firms' performances. We control for other inter-firm linkages. In the Internet Appendix B.1, we list data sources and provide information on variable construction for other measures of firms' overlap, such as product-market similarity, app user-base overlap, supply-chain linkages, geographic overlap, technology proximity (based on patent information), stock-market analyst overlap, etc.

Firms' product-design choices. To test a unique prediction of our model in Section 4, we measure firms' usage of functionality SDKs that reflect firms' product-design choices rather than data sharing. For each of the following functionality SDK categories—payment, security, customer

[19] Quarterly earnings growth rate is calculated as $2 \cdot (\textit{Net Income}_t - \textit{Net Income}_{t-1})/(\textit{Net Income}_t + \textit{Net Income}_{t-1})$, instead of $(\textit{Net Income}_t - \textit{Net Income}_{t-1})/\textit{Net Income}_{t-1}$ to smooth out volatility.

[20] There are ten 12-month non-overlapping windows, 7 windows before the introduction of ATT in April 2021 and 3 windows afterwards. 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.

support, review & feedback—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 (data-connected) firms, 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, which according to our model, affect the focal firm's product-design choices. These functionality SDKs are pre-built software components (just as the data-sharing SDKs are), which firms can integrate into their mobile apps. Apptopia tracks these app features by analyzing the installation packages of each new app version.

2.3 Descriptive statistics

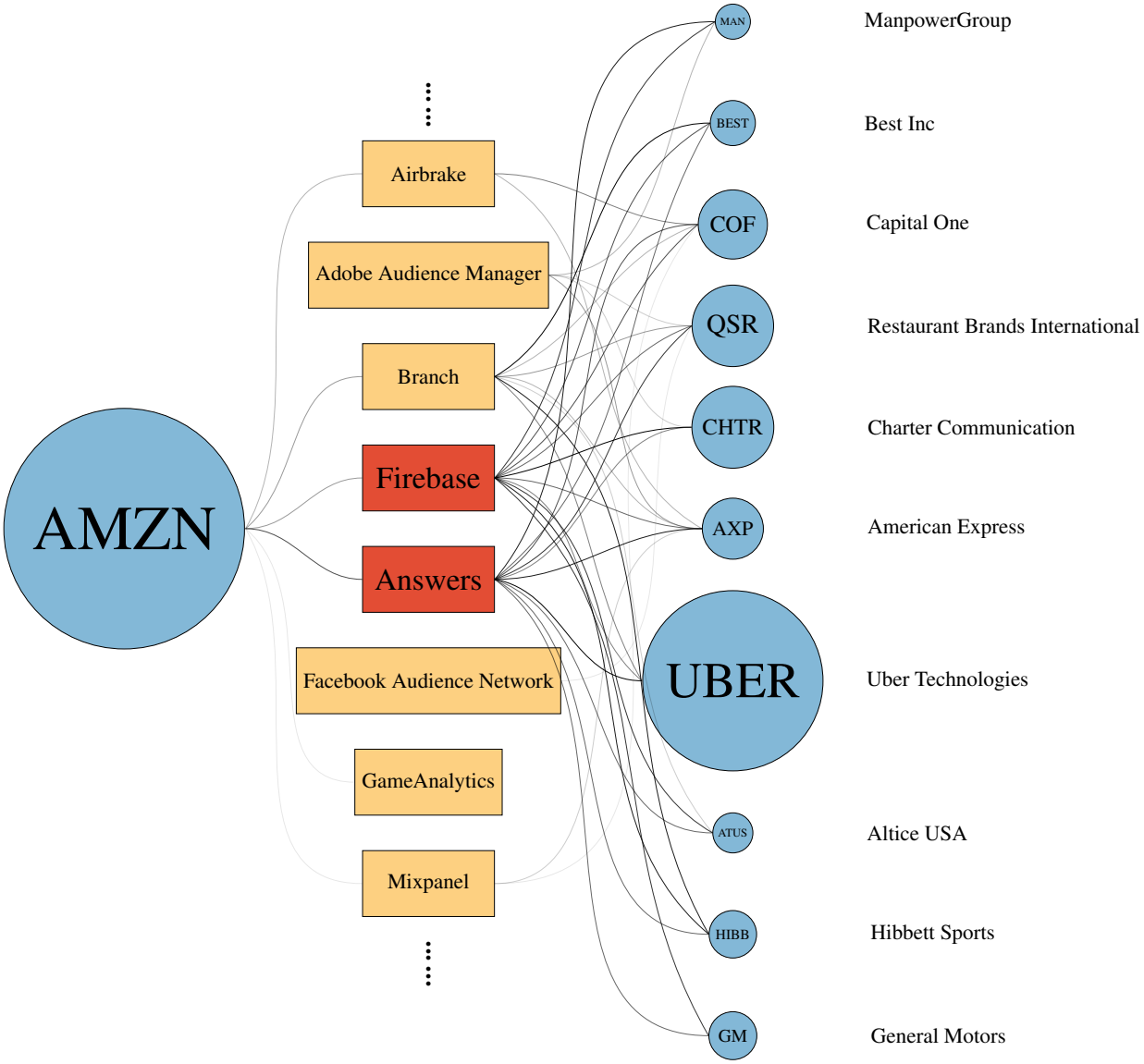
The data-sharing conduits. In Section 2.1, we use Amazon as an example of how a firm shares data with other firms (e.g., GM). In Figure 2, we map out the major data-sharing conduits that connect Amazon with other firms. Data-related SDKs that Amazon installed on its apps are shown as red or yellow rectangles. The red-colored SDKs are the most important as they connect Amazon with *all* the other firms in the figure. The thickness of lines connecting firms and SDKs represents the proportion of a firm's monthly active users (MAU) linked to an SDK.

Next, we map out the entire network structure of data-sharing conduits that is implied by our measure of data-connectedness. Figure 3 visualizes the network structure based on the Kamada-Kawai Algorithm that places the most well-connected firms centrally in the graph. The pairwise data connectedness takes the average value in 2020. Each node represents a firm, and the size of the node corresponds to the firm's size, measured by the square root of average MAU in 2020. For readability, we only include firm pairs with data connectedness greater than 0.7, which includes 649 unique firms (63% of firms in our sample). Edge color darkens proportionally with the strength of the data connection between two firms. We label the stock tickers of firms with on average more than 6.5 million MAU in 2020. Firms are grouped into different colors ("community") based on the Louvain Community Detection Algorithm, and in the legends below the graph, we label the most popular (widely installed) SDK within each community.[21]

Google (ticker: GOOGL) and Meta (ticker: META) appear to be the two most influential firms in the data network, evident from both their node sizes and central locations. Other telecom firms, such as AT&T (ticker: T) and Comcast (ticker: CMCSA), are also located near the center. Several firms from other industries, including Sony (ticker: SONY) from entertainment, DoorDash (ticker: DASH) from transportation, and McDonald's (ticker: MCD) from restaurants, also take central positions. Importantly, a firm's size does not always correlate with its network centrality.

[21]The algorithm partitions the network of nodes into community by maximizing modularity (the strength of community division), which measures the density of links inside communities compared to links between communities.

FIGURE 2: **Amazon's Data-Sharing Conduits and Peers**

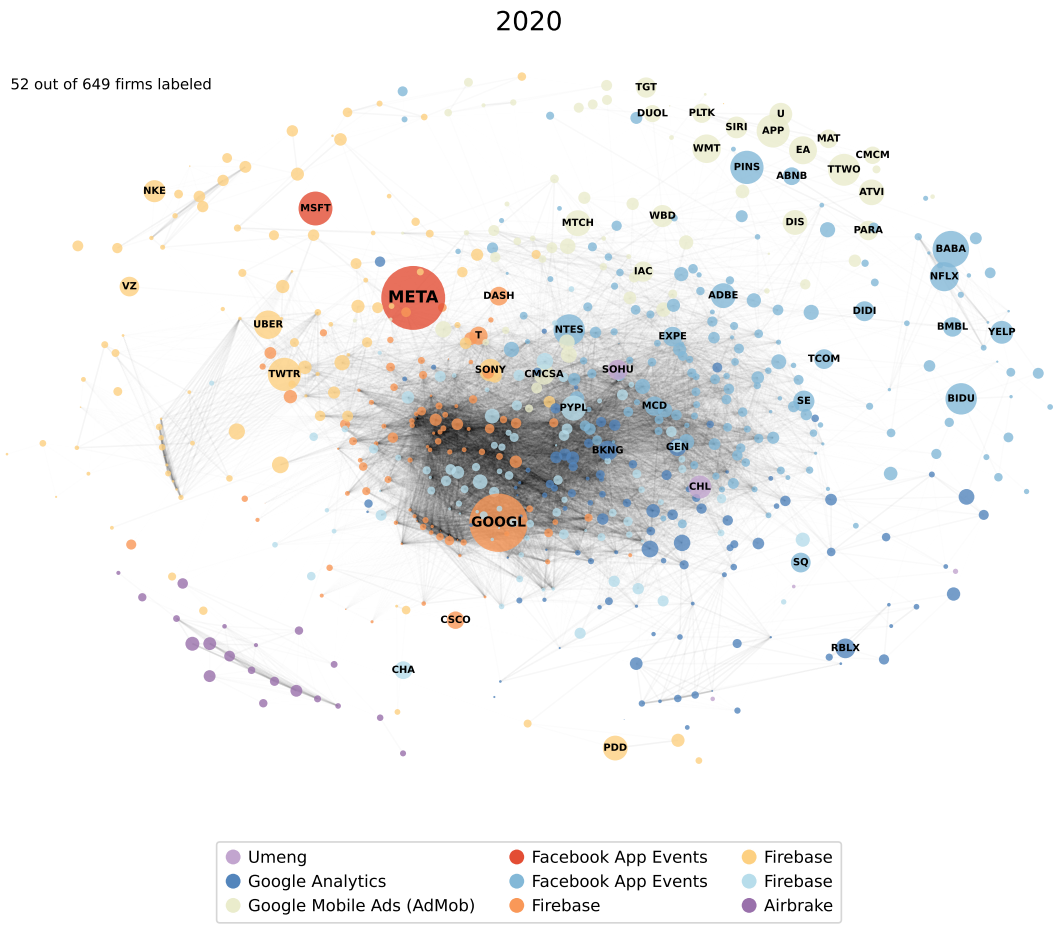


NOTE.— Figure 2 illustrates Amazon and the 10 firms most connected to it in the data space. Firms are depicted as blue circles, with the circle size corresponding to the firm's Monthly Active Users (MAU). Data-related SDKs are shown as red or yellow rectangles. The red-colored SDKs are those that connect Amazon with all the other firms in the figure. Lines connecting firms to SDKs indicate SDK usage, 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.

For example, large firms like Nike, Walmart, and Netflix remain on the periphery, while many relatively smaller firms occupy central positions in the network due to their interconnectedness.

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

FIGURE 3: **The Data-Sharing Network**



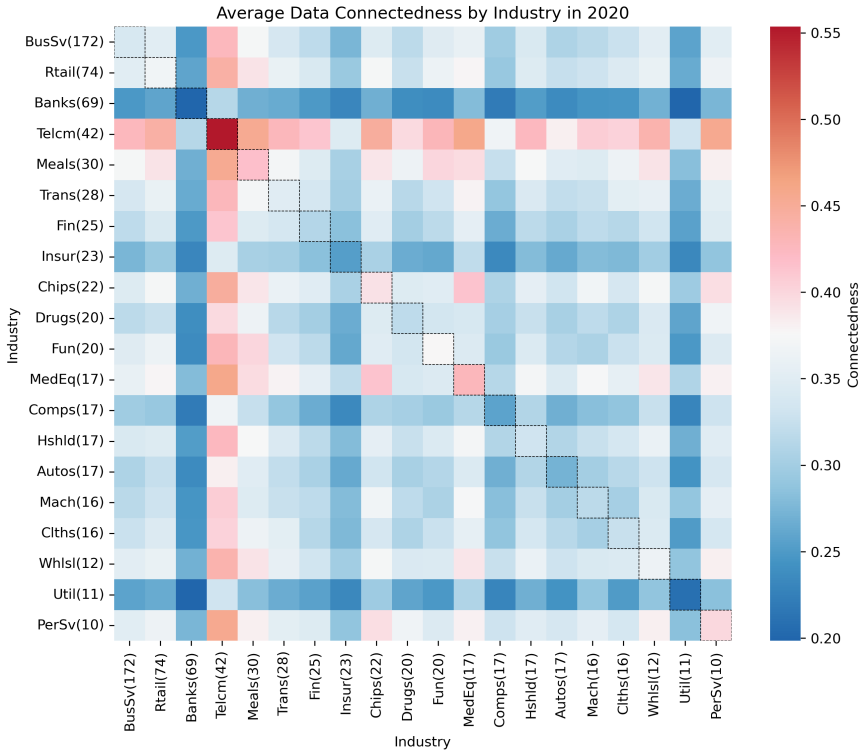
NOTE.— Figure 3 visualizes the network of firms connected in the data space using the Kamada-Kawai Algorithm and average data connectedness in 2020. Detailed explanation is provided in-text.

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.

Data sharing within and across industries. In the case study of Amazon (see Figure 2), it is apparent that firms across industries share data with one another. Next, we compare cross- and within-industry data sharing. One might expect that within an industry, firms would avoid adopting the same SDKs as they are reluctant to share data with competitors. However, a countervailing force exists: data from industry peers can be more informative about a firm's own customers than data from different industries. Empirically, these opposing forces appear to offset each other.

In Figure 4, firms within the same industry share data to a degree very similar to that observed between firms in different categories. The diagonal elements represent data connectedness

FIGURE 4: **Data Connectedness within and across Industries**



NOTE.—Figure 4 presents the average pairwise data connectedness across all the Fama-French 48 industry combinations in 2020, focusing on the 20 industries with the largest number of firms in our sample. Warmer (cooler) colors indicate higher (lower) levels of connectedness.

among firms within the same industry, while the off-diagonal elements capture cross-industry interconnectedness. Data sharing along the diagonal is neither systematically stronger nor weaker than that off the diagonal. One exception is the telecom industry, which exhibits stronger data connectedness with all other sectors. This is because telecom firms' apps typically install more SDKs.[22] Another interesting pattern is that heavily regulated industries, such as banks, insurance, and utilities, share less data with other industries. To demonstrate that our results are not specific to within- or cross-industry sharing, our baseline analysis in the main text considers all firm pairs, and in the Internet Appendix D, we only consider cross-industry firm pairs.

Summary statistics. Our measure of data connectedness for each firm pair is a cosine similarity score, which has an average of 0.171 and 90$^{\text{th}}$ percentile at 0.45. Note that data connectedness is stable over time. The average year-over-year Spearman correlation of our pairwise measure is 0.80, which is comparable to average correlation of 0.78 in the product space (Hoberg and Phillips,

[22]Internet Service Providers are the Internet gatekeepers and in a unique position to observe and capture rich cross-app and cross-device data—they install most of the popular SDKs to monetize from data mainly through advertisement.

TABLE 1: **Summary Statistics**

	mean	sd	p10	p25	p50	p75	p90	count
Pairwise connections								
data connectedness	0.171	0.19	0.00	0.00	0.11	0.29	0.45	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	0.247	0.33	-0.21	0.02	0.27	0.50	0.67	5,748,100
return - CAPM	0.034	0.33	-0.41	-0.21	0.03	0.27	0.47	5,645,298
return - DGTW	0.008	0.33	-0.42	-0.23	0.01	0.24	0.44	4,993,880

NOTE.—Table 1 reports the summary statistics on key variables. It lists all variables constructed at the firm-pair level. The first row reports statistics of data connectedness defined in (1). The other rows are for the outcome variables (i.e., firms' performance comovements). Downloads and DAU are measured in logs. The comovement of app and financial performances 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 2014Q3 to 2023Q2.

2010, 2016; Frésard, Hoberg, and Phillips, 2020) in our sample period. Our analysis thus treats the data-sharing conduits as given and focuses on its impact on firms' performances and decision-making. Firms' SDK installation decision lies beyond the scope of this paper. While we do not analyze changes in the data conduits, we verify that our results remain robust when we average our measure of data connectedness in different sample periods (see the Internet Appendix D).

We examine how data connectedness affects firms' correlations in app and financial performances. These correlations vary substantially across firm pairs. For example, the 10^{th} and 90^{th} percentile of log-downloads correlation are -0.67 and 0.73, respectively. For stock returns, there is stronger comovement in raw returns (0.247) compared to CAPM abnormal returns (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.[23] Table C.1 in the Internet Appendix reports summary statistics of all variables, including the other inter-firm linkages, firm characteristics, and firms' choices of (non-data) functionality SDKs that reflect their product design.

[23]The pairwise return correlations have more observations than the performance correlations: 10 per pair from rolling 12-month windows versus 2 per pair (one each for the pre- and post-ATT periods) as described in Section 2.

3 Data Sharing and Firm Performances

3.1 Performance comovement

Data is generated as a by-product of business operations (e.g., Bergemann et al., 2022; Jones and Tonetti, 2020; Farboodi and Veldkamp, 2021). When a firm expands its operation, it collects more data through increased interaction with customers. The new data is shared with its SDK-connected peers. Being more informed about customers allows the peer firms to grow. We now test the hypothesis that two data-connected firms exhibit performance comovement. We also examine the impact of ATT that allows customers to sever the link between their data and their universal identifiers (i.e., IDFA discussed in Section 2.1). By disrupting data aggregation based on universal IDs, ATT can weaken firms' performance comovement induced by data sharing.

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:

$$Corr_{ijt}^{Perf} = \alpha + \beta_1 \rho_{ij}^{data} + \beta_2 ATT_t \times \rho_{ij}^{data} + \beta_3' \rho_{ij}^{other} + \beta_4' ATT_t \times \rho_{ij}^{other} + \theta_{it} + \iota_{jt} + \varepsilon_{ijt}\,. \quad (2)$$

The correlation metrics, $Corr_{ijt}^{Perf}$, capture firm pairwise comovement across multiple dimensions, where *Perf* can be firms' 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 10 observations (7 before ATT and 3 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 firm linkages, we control for eight alternative linkages, denoted by ρ_{ij}^{other} (an 8-by-1 vector), which have been shown in prior literature to impact firms' comovement. These linkages include product-market overlap, supply chain relationships, technological proximity, common analyst coverage, etc. For all the firm linkages, including our measure of data connectedness, ρ_{ij}^{data}, we use the pre-shock (ATT) average values.[24]

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 negatively impacts firms' data collection from operations, thereby weakening the mechanism behind performance comovements between data-connected firms. To facilitate the interpretation of the coefficients and comparison

[24]In Appendix D, we consider alternative time windows for calculating average data connectedness, such as the average of ρ_{ijt}^{data} in 2020 (the year immediately preceding ATT) and post-ATT averages of ρ_{ijt}^{data}. We also use pre- and post-ATT averages of ρ_{ijt}^{data}, respectively, to explain firm performance comovement in the corresponding periods. All results are consistent with our baseline findings based on the pre-ATT average of ρ_{ijt}^{data}.

TABLE 2: **Comovement in Customer Engagement (App Performance)**

	downloads		DAU	
	(1)	(2)	(3)	(4)
data connectedness	0.026***	0.024***	0.026***	0.023***
	(7.56)	(7.25)	(6.78)	(6.26)
ATT × data connectedness	-0.025***	-0.025***	-0.024***	-0.023***
	(-6.63)	(-6.74)	(-5.81)	(-5.75)
Other linkages	N	Y	N	Y
ATT × other linkages	N	Y	N	Y
Observations	1,401,082	1,401,082	1,399,426	1,399,426
R-sq	0.066	0.068	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 pairwise 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.

across different types of firm linkages, we normalize ρ_{ij}^{data} and elements in ρ_{ij}^{other} to have a zero mean and a standard deviation equal to one. The results are reported in Table 2 to 4. In the Internet Appendix, Table A.1 to Table A.3 report the full results with estimates of β_3 and β_4.

Table 2 shows that higher data connectedness is associated with a significantly greater degree of app performance comovement (i.e., $\beta_1 > 0$). In Column 1 of Table 2, a one-standard-deviation increase in data connectedness leads to a 0.026-unit increase in the correlation of log(downloads) between two firms. Controlling for alternative types of firm linkages has little effect on the results in Column 2. We can compare the magnitude of these effects against the effects of other firm linkages. Based on Column 2 in Table A.1 in the Internet Appendix where we report the full results, the estimated coefficient of 0.024 for ρ_{ij}^{data} is 2.7 times larger than that for firms' product-market overlap, which is 0.009. We find similar results for the comovement in log(DAU).

This pattern is substantially weaker after ATT. Examining the ATT impact serves to validate that our measure of data connectedness indeed captures data-related linkages. If our measure were instead reflecting other economic relationships unrelated to data sharing, then ATT should have had a similar impact on other forms of firm linkages. The coefficients for the interaction terms between ATT and ρ_{ij}^{other}—the non-data linkages—are not significantly negative (see Table A.1).

In terms of magnitude, the coefficient of $ATT \times \rho_{ij}^{data}$ almost offsets the baseline effect. This does not imply that comovement driven by data sharing disappears. Our measure of data connectedness, ρ_{ij}^{data}, focuses on data sharing through SDKs in the iOS system where, before ATT,

TABLE 3: **Comovement in 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)
Other linkages	N	Y	N	Y
ATT × other linkages	N	Y	N	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.

data from different firms was merged by SDKs primarily based on IDFA (iOS app users' universal ID). ATT affected IDFA-based data merging. Other methods of data merging, for example, based on firms' proprietary customer IDs, persisted and may gain more adoption over time. In addition, ATT not only reduces the number of customers, whose data can be merged across firms based on IDFA, but also data representativeness. The iOS users who opted in IDFA-based tracking (between 4% to 18% in the first 12 months after ATT) may behave differently from the pre-ATT average iOS user. Such discontinuity in customer profiling may force firms to temporarily rely less on the SDK-transmitted data, thus weakening the performance comovement induced by 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. The smaller effect of data connectedness on earnings relative to that on app activities is intuitive, as firms have business activities and sources of profits unrelated to customers' app activities. This effect is again weakened by ATT. We obtain similar results with asset turnover as the measure of financial performance. For comparison, the effect of data connectedness on performance comovement is approximately 40% that of product (horizontal) overlap, as shown in Table A.2 in the Internet Appendix where we report the full results.

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

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***
	(6.52)	(3.64)	(6.71)	(4.35)	(6.25)	(3.76)
ATT × data connectedness	-0.002***	-0.002***	-0.003***	-0.003***	-0.002**	-0.002***
	(-3.00)	(-3.05)	(-3.86)	(-3.95)	(-2.53)	(-2.67)
Other linkages	N	Y	N	Y	N	Y
ATT × other linkages	N	Y	N	Y	N	Y
Observations	5,748,100	5,748,100	5,645,298	5,645,298	4,993,880	4,993,880
R-sq	0.442	0.449	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 monthly returns over rolling 12-month windows, relative to the introduction of ATT in April 2021. The return data spans from 2014Q3 to 2023Q2, 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. t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

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. After ATT, this effect 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 (see Table A.3 in the Internet Appendix), which captures firms' app similarity and therefore can be correlated with data connectedness.

Robustness and placebo tests. In the Internet Appendix D, we consider an array of alternative measures of data connectedness and regression specifications, Our results remain robust. First, our baseline measure of firms' data connectedness relies on the top 50 SDKs. For robustness, we verify that the results hold when restricting to the top 20 SDKs. Second, while the baseline measure treats all SDKs equally, our robustness tests allow for two sources of heterogeneity. First, we weight SDKs by their market share, giving greater influence to widely used SDKs. Second, we scale s_{iakt} (used in ρ_{ijt}^{data} in Equation (1)) by a measure of data-collection intensity, reflecting how actively each SDK acts as a data conduit. This adjustment accounts for the fact that firms with a similar number of app users may transmit different volumes of data. Finally, we consider averaging ρ_{ijt}^{data} over different time periods (our main results are based on the pre-ATT average values). We find similar results as the data conduits (SDK overlap) remain stable over time.

Moreover, we explore two alternative samples. First, we exclude firms that are themselves SDK providers, as these firms may play confounding roles in the data economy. Second, we

exclude firm pairs with strong product-market overlap and focus on firms' comovement induced by cross-industry data sharing. This addresses potential concerns about firms' willingness to share data with their competitors (see our discussion of Figure 4 in Section 2.1). In both cases, our estimates remain very similar to the baseline. We also show that the results also remain highly significant when we cluster standard errors at alternative levels.

Finally, in Table A.4 of the Internet Appendix, we conduct a placebo test. Using the same approach as in Section 2.2, we measure firms' SDK overlap based on SDKs unrelated to data sharing, and show that neither this alternative SDK overlap nor its interaction with the ATT indicator explains firms' performance comovement. As a reminder, our main results have already validated that our data connectedness measure captures data sharing: its interaction with the post-ATT dummy shows a dampening effect from ATT, while, crucially, this dampening effect of ATT is absent for other inter-firm linkages. This placebo test provides further support for our measure.

3.2 Cyberattack ripple effects

Performance comovement emerges from shock propagation across firms—that is, a shock that originates from one firm can impact another firm through data connectedness, thus generating comovement. Next, we characterize such shock propagation. Specifically, we examine the propagation of cyberattacks—a particularly salient type of shock in the data economy.

We collect major cyberattack events using data from Advisen that covers more than 90,000 cyber events between 2000 and 2023 from publicly verifiable sources. We identify 22 major cyber events involving at least 10 million exposed records (listed in Table B.1 in the Internet Appendix).

For a local firm involved in a cyber event τ, we define a peer firm i's exposure as:

$$\text{Exposure}_{i\tau} = \frac{\sum_P \rho_{i\tau}^{\text{data},P} DAU_\tau^P}{\sum_P \sum_j \rho_{ij}^{\text{data},P} DAU_j^P}$$

where j represents a firm connected to firm i via data sharing, and P represents platforms, $P \in \{\text{iOS, Android}\}$, and $\rho_{i\tau}^{\text{data},P}$ is data connectedness between firm i and the focal firm involved in event τ, as defined in Equation (1).[25] A higher value of $\text{Exposure}_{i\tau}$ implies that the impacted firm's data is important for firm i relative to all the other connected firms j. A firm i is considered to be highly exposed event if $\text{Exposure}_{i\tau} \geqslant 0.01$, corresponding to the 75$^{\text{th}}$ percentile of the exposure distribution. Firms with $\text{Exposure}_{i\tau} < 0.01$ are considered control firms. We assign an indicator

[25] Note that in the regression analysis of firms' performance comovement, our data connectedness measure is based on ρ_{ij}^{iOS}, as ATT affects data collection and sharing only within the iOS ecosystem. In contrast, for the analysis of cyberattack impacts, we consider data connections between firms across both iOS and Android systems, since unlike ATT, which is an iOS-specific shock, cyberattacks affect firms regardless of their customers' operating systems and can propagate through both the iOS and Android data networks.

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

	log(downloads)		log(DAU)	
	(1)	(2)	(3)	(4)
cyber event × high exposure	−0.116***	−0.107***	−0.148***	−0.137***
	(−6.60)	(−5.31)	(−7.28)	(−5.75)
cyber event × other overlaps	N	Y	N	Y
Observations	81,373	81,373	81,373	81,373
R-sq	0.955	0.955	0.957	0.957

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 firm k is considered an important peer if $\text{Exposure}_{ik} > 0.01$, corresponding to the 75$^{\text{th}}$ 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.

variable, *high exposure*$_{i\tau}$, which takes the value of 1 for highly exposed firms and 0 otherwise.

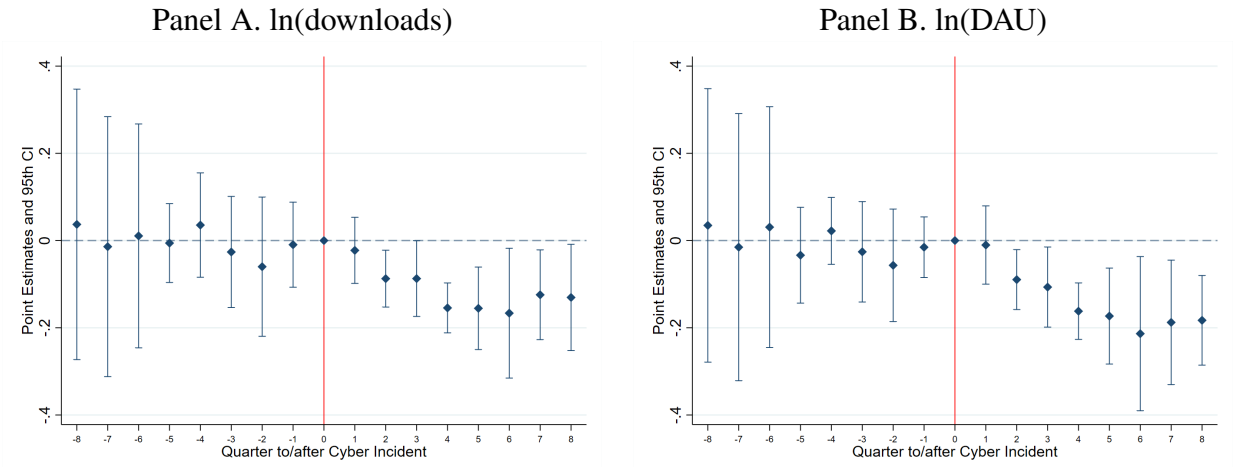
We estimate the following equation for the 22 cyber events using a stacked difference-in-differences approach:

$$Y_{i\tau t} = \alpha + \beta_1 cyber_\tau \times high\ exposure_{i\tau} + \beta_2 cyber_\tau \times \rho_{i\tau}^{other} + \beta_3' \mathbf{C_{it}} + \theta_{i\tau} + \iota_{\tau t} + \varepsilon_{i\tau t}, \quad (3)$$

where τ indexes events, i indexes firms, and t indexes the quarter relative to each event. We include 8 quarters before and 8 quarters after the event. We consider two outcomes for Y_{it}: app downloads and DAU (daily active users) at firm-quarter level, both in logarithm. *cyber*$_\tau$ is an indicator variable that takes the value of 1 after the cyberattack. $\rho_{i\tau}^{other}$ includes non-data linkages between i and the focal firm in the event τ. In all specifications, we include the following firm-level controls, $\mathbf{C_{it}}$: firm size (log of assets), long-term debt, tangible assets, cash, R&D expenditure, and capital expenditures, with the latter five normalized by total assets. Additionally, we control for firm-by-event fixed effects ($\theta_{i\tau}$) to isolate event-specific effects, as a given firm may be exposed to multiple events. We also include event-specific quarter fixed effects ($\iota_{\tau t}$) to absorb common shocks affecting all firms around the 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 ln(downloads) and the last two columns on ln(DAU). In all columns, β_1 is negative and statistically significant at the 1% level. The magnitude of the cross-firm spillover effect of these major cyberattacks is economically large. Specifically, relative to the control firms, the highly exposed firms experience a 11.6% drop in quarterly downloads (Column 1) and a 14.8% drop in DAU (Column 3). Controlling for the interaction of *cyber*$_\tau$ with other firm linkages, these point estimates

FIGURE 5: **Cyberattack Spillover Effects on App Performances**



NOTE.–Figure 5 shows the cross-firm spillover effects of 22 major cyber events. The dynamic DiD coefficients are obtained from estimating the dynamic version of Equation (3). A firm k is considered an important peer if Exposure$_{ik} > 0.01$, corresponding to the 75$^{\text{th}}$ percentile of the exposure distribution. Firms with 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.

drop slightly to 10.7% and 13.7%, 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. The treatment effect begins to emerge immediately after the cyber event occurs, continuing through the fifth quarter post-event. Finally, it stabilizes after six quarters with a slight upward reversal.

Furthermore, the spillover effect is not driven by cybersecurity vulnerability in firms' shared software design. In Table A.6 in the Internet Appendix, we show that the spillover effect disappears when we replace the data-sharing SDK overlap with overlap in SDKs that are unrelated to data sharing but reflect firms' overall app design, such as those that enable media display and within-app messaging. This placebo test confirms that the spillover effects emerge from data sharing.

Our findings provide the first evidence of how shocks to firms spill over to their peers through the data-sharing network. Such shock propagation applies to both firms' app performances and stock-market valuation. In the Internet Appendix, we conduct event studies on data-connected firms' stock performances surrounding cyberattacks on the focal firm.[26] Firms that are data-connected to the focal firms experience a 3-4% decline in cumulative (CAPM-based) abnormal returns in the month. Using other benchmark models for computing abnormal returns, such as the Fama-French three factor model and DGTW, yields qualitatively similar results.

[26]In Figure A.3, Panel A and Panel B use the incident dates and notice dates, respectively, as the event dates. Market participants were likely to be made aware of those events between the incident dates when cyberattacks happened and notice dates when firms made announcement typically following preliminary investigations.

4 A Network Model of Data Economy

We develop a model of data economy where firms are interconnected through data sharing. Data improves firms' understanding of customer preferences and thereby stimulates customer engagement and contributes to revenue generation. The model captures our empirical findings on firms' performance comovements and cross-firm shock propagation in Section 3. We develop a firm-valuation framework that captures data spillover effects of higher orders and over multiple time horizons and, using this framework, we identify systemically important firms in the data economy.

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 raw 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

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

where the terms in the bracket are a stochastic growth rate, and $z_{i,t}$ is a standard Brownian motion that is independent across firms.[27] Data depreciates when the last term is negative. Faster customer turnover is associated with faster data depreciation, as data is typically used for customer profiling.

The first term on the right side of (4), $y_{i,t}dt$, captures the idea that data is generated by customer activities, where θ (> 0) represents data collection efficiency. An e-commerce platform collects data on consumers by observing their transactions and search activities. For a retail company, promotions of products generate interactions with customers, and the feedback from customers is informative of their preferences. Software companies learn about users' preferences through their usage patterns. Our empirical setting covers many industries. Accordingly, we set up our model to be sufficiently generic to highlight the commonality among these firms.

In the law of motion (4), customer engagement, $y_{i,t}$, generates new data. Data in turn helps firms stimulate customer activities by making firms more informed—that is, data contributes to $y_{i,t}$.[28] This force leads to the self-perpetuating growth of data in the spirit of Farboodi and Veldkamp (2021). Next, we discuss how $y_{i,t}$ is determined and its dependence on data.

In our model, data has two forms, the raw data collected at the firm level, $\delta_{i,t}$, and the processed and aggregated data, $D_{i,t}$, that corresponds to the comprehensive customer profiles that

[27] A negative shock may reflect a direct loss of data, for example, due to cyberattack. Florackis, Louca, Michaely, and Weber (2022) measure cybersecurity risk from corporate disclosure. In contrast, following a positive shock, the firm gains more information on its customers, for example, through an increase in product reviews.

[28] 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 time spent on the product.

firms obtain from the SDKs in our empirical setting:

$$D_{i,t} = \sum_{j=1}^{N} \gamma_{ij} \xi \delta_{j,t} = \{\xi \Gamma \bar{\delta}_t\}_i, \tag{5}$$

where γ_{ij} represents the data conduit from firm j to i facilitated by the shared SDKs, and $\xi \in [0,\ 1]$ is the fraction of raw data linked to universal IDs (e.g., IDFA). Note that consistent with our empirical findings, the data conduits is stable over time, so γ_{ij} does not have a time subscript. As discussed in Section 2.1, universal IDs allow data from different firms to be merged into customer profiles. In iOS systems, $\xi = 1$ before ATT. After ATT, $\xi < 1$ because iOS users may disable IDFA-based data tracking. Let $\bar{\delta}_t$ denote the column vector of all firms' raw data: $\bar{\delta}_t = [\delta_{1,t},\ ...,\ \delta_{N,t}]^\top$. We introduce the notation Γ where the ij-th element is γ_{ij} and the row sums are equal to one. Multiplying Γ on $\bar{\delta}_t$ we obtain a vector of SDK-transmitted data, and the i-th element of $\xi\Gamma\bar{\delta}_t$, denoted by $\{\xi\Gamma\bar{\delta}_t\}_i$, is firm i, is firm i's information on its customers, $D_{i,t}$.

The processed information, $D_{i,t}$, incorporates firm i's own raw data and data from other firms. At time t, customers' preferences, denoted by $\widetilde{\eta}_{i,t}$, is drawn from a Gaussian distribution, $\mathcal{N}(0,\ \sigma_\eta^2)$, and firm i receives a Gaussian signal, denoted by $h_{i,t}$, with a zero mean. The firm forms a forecast, $a(h_{i,t})$, that maximizes precision, or equivalently, minimizes the mean-squared error:

$$\lambda_{i,t}^2 := \left(\min_{a(h_{i,t})} \mathbb{E}\left[(a(h_{i,t}) - \widetilde{\eta}_{i,t})^2 \right] \right)^{-1}, \tag{6}$$

where $\lambda_{i,t}^2$ is the precision. The firm faces an information capacity constraint as in Sims (2003):

$$\mathcal{I}\left(\widetilde{\eta}; h_{i,t}\right) = \mathcal{H}\left(\widetilde{\eta}\right) - \mathcal{H}\left(\widetilde{\eta}|h_{i,t}\right) \leq \ln\left(\kappa D_{i,t}\right), \tag{7}$$

where we use $\mathcal{I}\left(\widetilde{\eta}; h_{i,t}\right)$ to denote the mutual information and $\mathcal{H}\left(\widetilde{\eta}\right)$ and $\mathcal{H}\left(\widetilde{\eta}|h_{i,t}\right)$, respectively, to denote the unconditional entropy and h-conditional entropy of $\widetilde{\eta}$. The parameter, $\kappa > 0$, scales $D_{i,t}$ into information units in logarithm. We assume $\kappa D_{i,t} > 1$. The constraint (7) restricts the informativeness of firm i's signal $h_{i,t}$. This problem has a classic solution: $a(h_{i,t}) = \mathbb{E}\left[\widetilde{\eta}_{i,t}|h_{i,t}\right]$ with the mean-squared error given by $\left[\sigma_\eta/(\kappa D_{i,t})\right]^2$, which then implies $\lambda_{i,t} = \kappa D_{i,t}/\sigma_\eta$.

We specify customer engagement, $y_{i,t}$, as follows:

$$y_{i,t} = \alpha \lambda_{i,t} + x_{i,t}, \tag{8}$$

where $\alpha > 0$, implying the self-perpetuating growth of data: firm i's raw data $\delta_{i,t}$ enters $D_{i,t}$ in (5), which in turn contributes to $y_{i,t}$ through the forecasting precision, $\lambda_{i,t}$, in (8), and $y_{i,t}$ generates more data $d\delta_{i,t}$ in (4). Different from Farboodi and Veldkamp (2021), our model features data

spillover and interconnected growth, as other firms' data affects firm i's data growth through the forecasting precision, $\lambda_{i,t}$, and then $y_{i,t}$, and vice versa.

Beyond a data-enabled level of customer engagement in the first term in (8), the firm can enhance customer engagement by an extra amount, $x_{i,t}$, through a product-design choice: a higher $x_{i,t}$ makes its product more suitable for generating customer engagement rather than cash flows (monetization). Next, we introduce the trade-off in firm i's choice of $x_{i,t}$.

Cash flows. Customer engagement does not equate to cash-flow generation but they are connected in two ways. First, as shown in (4), $\delta_{i,t}$ represents firm i's raw data stock collected from the history of customer engagement, $y_{i,t}$, that has accumulated over time subject to stochastic growth or depreciation due to customer turnover. This history of interaction with customers forms firm i's customer capital that generates cash flow $\zeta\delta_{i,t}$ at time t.[29] Note that customer capital arises from the *cumulative* customer engagement, $\delta_{i,t}$, rather than the current level of engagement, $y_{i,t}$.

The second connection (or more precisely, tension) between customer engagement and cash flows is that when firm i prioritizes engagement in product design (i.e., it increases $x_{i,t}$ in (8)), it generates less cash flows. For example, an app design focused on engagement may offer rich free features to maximize customers' time spent in the app. In contrast, a design that prioritizes monetization can increase revenues by moving more features behind the paywall at the expense of customer engagement. The impact of $x_{i,t}$ on firm i's cash flow is captured by a function $g\left(x_{i,t}, \lambda_{i,t}\right)$ with $g_x\left(x_{i,t}, \lambda_{i,t}\right) < 0$. Intuitively, we also assume that when firm i is more informed about its customers, the negative cash-flow impact of prioritizing customer engagement over monetization is mitigated, i.e., $g_{x\lambda}\left(x_{i,t}, \lambda_{i,t}\right) > 0$. For example, if a software company has better knowledge of its customers, it can prioritize engagement by offering many free features without reducing cash flows, as it can more effectively target high-purchase-intent users with its premium offerings.

In summary, firm i's cash flows are given by

$$F_{i,t} = \zeta\delta_{i,t} + g\left(x_{i,t}, \lambda_{i,t}\right). \tag{9}$$

Let ρ denote the discount rate. Firm i's valuation (i.e., its value function at $t = 0$) is given by

$$V^i\left(\delta_{i,0}, \{\delta_{j,0}\}_{j\neq i}\right) = \max_{\{x_{i,t}\}_{t=0}^{\infty}} \mathbb{E}\int_{t=0}^{+\infty} e^{-\rho t} F_{i,t} dt. \tag{10}$$

The state variables include firm i's and other firms' raw data that all contribute to firm i's information on its customers, $D_{i,t}$. In summary, $D_{i,t}$ relaxes the constraint (7) on forecasting customers' preferences and thereby contributes to cash-flow generation in two ways. First, a better knowledge

[29] Following Gourio and Rudanko (2014a,b), we have cash flows linear in customer capital, consistent with evidence in Einav, Klenow, Levin, and Murciano-Goroff (2021) that firms' revenues largely scale with the number of customers.

of customers (i.e., a higher $\lambda_{i,t}$) increases customer engagement (see (8)), and the accumulated interactions with customers ("customer capital") generate the first component of cash flow. Second, $\lambda_{i,t}$ affects the trade-off between customer engagement and monetization in product design.[30]

We follow the literature that models data as a by-product of firms' productive activities and as a key input in production (e.g., Bergemann et al., 2022; Farboodi and Veldkamp, 2021).[31] Our model extends the framework in three aspects. First, we separate productive activities in two related areas, customer engagement and cash-flow generation. The former generates data and, over time, allows customer capital to accumulate, indirectly benefiting the latter in the long run. Second, we introduce a contemporaneous trade-off between these two areas of productive activities. Third, by improving firms' forecasting precision (i.e., knowledge on customers' random preferences), data contributes positively to both areas.[32] Separating the two areas of productive activities brings out a meaningful intertemporal trade-off: increasing $x_{i,t}$ reduces current profits but, by enhancing customer engagement, $y_{i,t}$, it increases $\delta_{i,t}$ that in turn boosts future profits.

Therefore, the choice of $x_{i,t}$ is akin to a (data-) investment decision. 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.[33] 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. Figure 6 illustrates the three blocks of our model, data dynamics, product-design choice, and firm valuation that we solve in Section 4.2.

Discussion: the characteristics of data assets. Our model captures the three features of data as a productive asset. First, as previously discussed, data is a by-product of firms' operations and in turn contributes to firms' operations. 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 deviates from externality at the individual level—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. We will show that in our model,

[30]Firms may use data to improve customers' willingness to pay (Ichihashi, 2020). The potential harmful impact of price discrimination is beyond the scope of this paper, as our focus is on firm dynamics rather than consumer welfare.

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

[32]In models that do not separate productive activities in two areas, data, often modeled through forecasting precision, typically contributes directly to firms' productivity (e.g., Farboodi and Veldkamp, 2021).

[33]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 financing or intensifying product-market competition that other firms face (e.g., Asriyan et al., 2024).

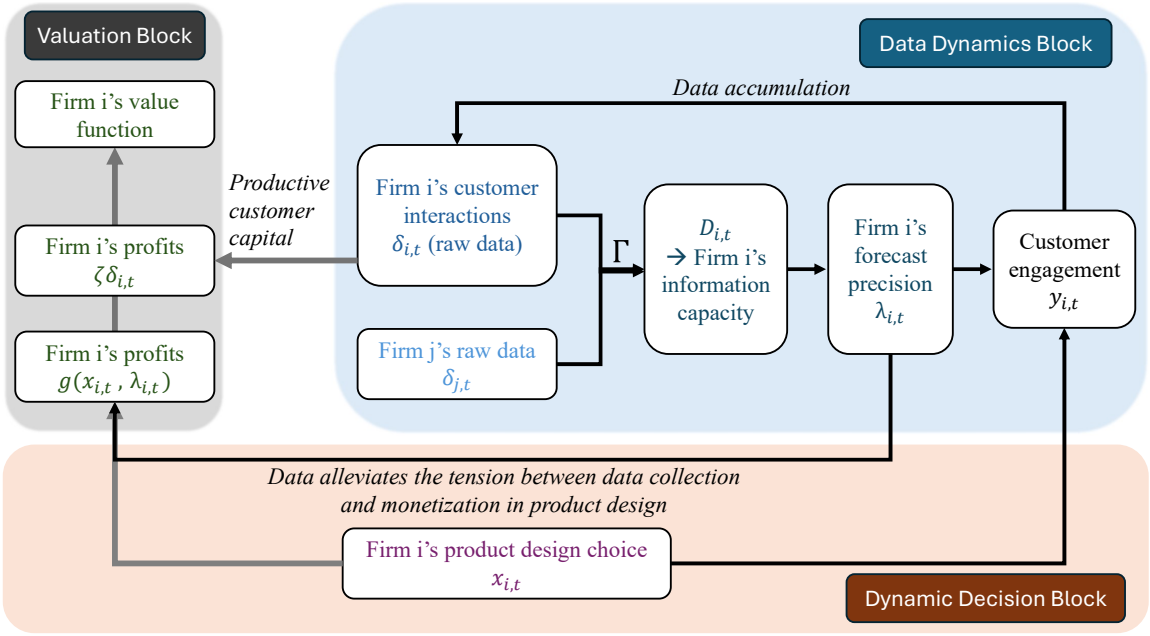


FIGURE 6: **A Summary of Model Structure**

firms under-invest in data accumulation under positive data externalities (and thus monetize excessively rather than prioritizing customer engagement). 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 features distinguishes 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, R&D is typically not a by-product of firms' interactions with customers.[34] In contrast, data is generated as a 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 raw data stock, $\delta_{i,t}$, is a state variable, and, due to the dependence on other firms' raw data via $D_{i,t}$, the other firms' $\delta_{j,t}$ ($j \neq i$) are also state variables for firm i. Next we analyze the optimal choice of $x_{i,t}$ through the following Hamilton-Jacobi-Bellman (HJB) equation for the

[34] R&D is not a by-product of customer engagement. It often requires deliberate, resource-intensive investment, separate from the normal production process (Corrado, Hulten, and Sichel (2005); Hall, Mairesse, and Mohnen (2010)).

value function of firm i at time t, denoted by $V^i(\bar{\delta}_t) = V^i\left(\delta_{i,t}, \{\delta_{j,t}\}_{j\neq i}\right)$:

$$\rho V^i\left(\delta_{i,t}, \{\delta_{j,t}\}_{j\neq i}\right) = \max_{x_{i,t}} \zeta\delta_{i,t} + g\left(x_{i,t}, \lambda_{i,t}\right) + V^i_{\delta_{i,t}}\left[\theta(\alpha\lambda_{i,t} + x_{i,t}) + \mu_\delta\delta_{i,t}\right] + \frac{1}{2}V^i_{\delta_{i,t}\delta_{i,t}}\delta^2_{i,t}\sigma^2_{i,\delta} + \sum_{j\neq i}\left[V^i_{\delta_{j,t}}\left[\theta(\alpha\lambda_{j,t} + x_{j,t}) + \mu_{j,\delta}\delta_{j,t}\right] + \frac{1}{2}V^i_{\delta_{j,t}\delta_{j,t}}\delta^2_{j,t}\sigma^2_{j,\delta}\right]. \tag{11}$$

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 (8), i.e., $y_{i,t} = \alpha\lambda_{i,t} + x_{i,t}$, where, as previously discussed, $\lambda_{i,t} = \kappa D_{i,t}/\sigma_\eta$ and $D_{i,t}$ is given by (5). 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 under $g_x(x_{i,t}, \lambda_{i,t}) < 0$, 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*

$$-g_x(x_{i,t}, \lambda_{i,t}) = V^i_{\delta_{i,t}}\theta\,. \tag{12}$$

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.

Data functions as a form of productive capital, with a firm's product-design choices mirroring investment decisions.[35] 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.[36] A key distinction from the traditional investment models is that firms' choices of $x_{i,t}$ are interconnected through data sharing. Other firms' data, $\delta_{j,t}$, enters firm i's optimality condition (12) via firm i's information on its customers, $D_{i,t}$ and the precision of its forecast of customer preferences, $\lambda_{i,t}$.

To sharpen the intuition, we introduce the following functional form $g(x_{i,t}, \lambda_{i,t}) = \zeta\log(\phi_\lambda \lambda_{i,t} - \phi_1 x_{i,t})$ that satisfy the properties of $g(x_{i,t}, \lambda_{i,t})$ discussed in Section 4.1 (i.e., $g_x\left(x_{i,t}, \lambda_{i,t}\right) < 0$ and $g_{xx}\left(x_{i,t}, \lambda_{i,t}\right) < 0$). We define $\phi_0 = \phi_\lambda\kappa/\sigma_\eta$ so that given $\lambda_{i,t} = \kappa D_{i,t}/\sigma_\eta$, we have $g(x_{i,t}, \lambda_{i,t}) = \zeta\log\left(\phi_0 D_{i,t} - \phi_1 x_{i,t}\right)$, which directly depends on $D_{i,t}$. Under this functional form,

[35] 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, Mihet, Philippon, and Veldkamp (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 sufficiently large such that the marginal impact on current profits is negative. The firm accepts the negative impact on profits because the marginal value of data acquired through experimentation is positive. We explicitly focus on the region where such an explicit trade-off between current profits and data acquisition emerges.

[36] In our model, firms adjust product design while facing an intertemporal trade-off between current cash flows and future cash flows that benefit from data accumulation. In Chen (2024), firms face a different trade-off: improving product quality is costly but earns customers' attention that can be monetized through selling advertisements.

we obtain closed-form solutions of the value function. Below we define the *users' cost of data*:

$$\hat{\rho} = \rho - \mu_\delta. \tag{13}$$

In (4), a negative μ_δ represents data depreciation—stale information on customer behavior is less informative due to customer turnover. Therefore, $\hat{\rho}$ resembles the *user's cost of capital* in investment theory, i.e., the sum of required rate of return on capital (discount rate) and depreciation rate. The next proposition summarizes the solution of value function. Appendix E provides the proof.

Proposition 2 (Network-Augmented Gordon Growth Formula) *Under the parameter condition,*

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

firm i's value at time t is given by

$$V_{i,t} = \zeta \left\{ (\hat{\rho}\mathbf{I} - \beta\Gamma)^{-1} \bar{\delta}_t \right\}_i + v_{i,0}, \tag{15}$$

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—that is, when $\Gamma = \mathbf{I}$ ($\gamma_{ii} = 1$ and $\gamma_{ij} = 0$ for $j \neq i$)—each firm relies on its own data, and firm i's valuation reduces to the standard Gordon growth formula:

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

where ζ is the cash flow per unit of $\delta_{i,t}$, and $\hat{\rho}$ and β play the roles of discount rate and growth rate, respectively. Under data sharing, the firm-level growth is replaced with "communal growth" of the whole data economy, which in turn depends on Γ, the data-conduit matrix.

The composite parameter for growth, β, reflects a sequence of events, from data collection (ξ) to customer engagement directly driven by data ($\alpha\kappa/\sigma_\eta$) and customer engagement from product design choice (ϕ_0/ϕ_1), and then from customer engagement to firm growth (θ). Below we explain each step. All firms' raw data is recorded in the vector $\bar{\delta}_t$, where ξ fraction is linked to universal IDs. Firms' information on customers is given by $D_t = \xi\Gamma\bar{\delta}_t$, where the data conduits, Γ, allow raw data to be transferred and merged across firms. When forecasting customer preferences, firm i's precision, $\lambda_{i,t}$, depends on its information on customers, $D_{i,t}$ (the i-th element of D_t), through κ/σ_η, and then, $\lambda_{i,t}$ drives customer engagement, $y_{i,t}$, via $\alpha\lambda_{i,t}$ (see (8)). Customer engagement, $y_{i,t}$, and firm growth also depend on $x_{i,t}$. Increasing $x_{i,t}$ (prioritizing engagement over monetization) reduces cash flows. When ϕ_0 is higher, firm i's information on customers mitigates this negative impact, and a lower ϕ_1 also reduces this negative impact—that is, overall, a high ratio

of ϕ_0/ϕ_1 allows firm i to design product that stimulates customer engagement without sacrificing current profits too much. Finally, given customer engagement, $y_{i,t}$, firm i grows by accumulating raw data and customer capital—that is, $d\delta_{i,t}$ depends on $y_{i,t}$ through the parameter θ (see (4)).

The data-conduit matrix, Γ, appears in $\zeta\left\{(\hat{\rho}\mathbf{I}-\beta\Gamma)^{-1}\bar{\delta}_t\right\}_i$. This term can be decomposed as follows to show the rounds of data propagation through the data-conduit network:

$$\zeta\left\{(\hat{\rho}\mathbf{I}-\beta\Gamma)^{-1}\bar{\delta}_t\right\}_i=\frac{\zeta}{\hat{\rho}}\left\{\left(\mathbf{I}-\frac{\beta}{\hat{\rho}}\Gamma\right)^{-1}\bar{\delta}_t\right\}_i=\frac{\zeta}{\hat{\rho}}\left\{\sum_{k=0}^{\infty}\left(\frac{\beta}{\hat{\rho}}\Gamma\right)^k\bar{\delta}_t\right\}_i. \tag{16}$$

The first component, $\zeta/\hat{\rho}$, is the value of a cash-flow perpetuity without growth. The growth operator, $\sum_{k=0}^{\infty}\left(\frac{\beta}{\hat{\rho}}\Gamma\right)^k$, is applied to the data vector $\bar{\delta}_t$, incorporating direct and indirect network effects. Firm i is affected by a connected firm's data through its precision in forecasting customer preferences, which in turn influences its product-design decision, customer engagement, data collection, and growth. The variation in firm i's data then affects other firms, resulting in second-order spillover effects. The infinite sum accounts for all the direct and indirect channels of network propagation over time with $\beta/\hat{\rho}$ being the attenuation factor. The convergence of the geometric sequence is guaranteed by $\beta/\hat{\rho}<1$ under the condition (14).[37] The spillover effects weaken as the distance increases: k-degree network propagation, Γ^k, is "discounted" by $(\beta/\hat{\rho})^k$. The strength of network propagation depends on the aforementioned channels of data-driven growth, with each step captured by a component of β as previously discussed. The next corollary highlights the impact of ATT on firm growth and network propagation of data by reducing ξ.

Corollary 1 (ATT Impact) *ATT reduces ξ, the fraction of data linked to universal IDs, and thereby reduces β, the growth of data economy, and weakens the network effects of data sharing.*

Our valuation metrics naturally encompass the impact of data variation at one firm on all other firms across multiple time horizons and network spillover effects of first, second, and all the higher-order degrees. 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.

Motivated by our findings in Section 3, the model setup directly captures the comovement in firms' app performances (customer engagement, $y_{i,t}$) and cash flows due to data sharing. In addition, after solving firms' valuations, we compute the stock return, $dR_{i,t}=dV_{i,t}/V_{i,t}$.[38] The

[37] As a reminder, the data-conduit matrix, Γ, is right-stochastic so its largest eigenvalue is one.

[38] Our valuation exercise and return calculation take the perspective of an investor with perfect information. We

next corollary shows that data sharing induces return comovement and ATT weakens such return comovement in line with our empirical findings in Section 3. Appendix E provides the proof.

Corollary 2 (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 $\frac{\partial^2\rho_{r,i,j}}{\partial\xi\partial\gamma_{i,j}} > 0$.*

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 as in our model. Therefore, the model takes the entire data-sharing network 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 ζ by regressing firms' revenues on DAU to obtain $\zeta = 18$.[39] A closely related concept, average revenue per user (ARPU), is a common metric used in marketing research to evaluate the conversion of customer capital to cash flows.

We pin down a subset of parameters with estimates from prior studies. 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). We set a quarterly depreciation rate of 7.5% ($\mu = -0.075$), consistent with the 30% annual amortization schedule applied to data assets under the current accounting standards.[40]

The last two parameters, β and ξ, are calibrated to match six moments from our empirical findings in Section 3, as discussed below. Note that α, θ, ϕ_0, and ϕ_1, form a composite parameter β. Model simulation requires β as an input but does not require the values of these four parameters separately. We set β to 0.09. Moreover, we set ξ, the fraction of data linked to universal IDs, to one before ATT and 0.7 after ATT.[41] We summarize the calibrated parameters in Panel B of Table 6.

emphasize return comovement due to firms' interconnected fundamentals (networked data asset). This is different from Veldkamp (2006) that emphasizes comovements in asset prices due to market participants' information choices.

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

[40] Farboodi and Veldkamp (2021) point out that accounting rules depreciate data like software at 30% annually, but the actual depreciation rate of data may depend on whether the forecasting targets are static (e.g., consumers' tastes) or something more ephemeral like stock trading order flows. Abis and Veldkamp (2023) use a 3% monthly depreciation rate. The intangible capital literature reports similar figures. For example, Ewens, Peters, and Wang (2024) estimate 33% for knowledge capital and 28% for organizational capital. Customer churn rates, estimated to be around 30-35% annually, provide another justification, see He, Mostrom, and Sufi (2024) and Baker, Baugh, and Sammon (2023).

[41] Although around 82% of iOS users opted out of IDFA tracking as of April 2022, ATT's impact is likely smaller.

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

A. Moments		
	Data	Model
Regression coefficient of		
– App performance comovement on data connectedness	0.023	0.012
– Financial performance comovement on data connectedness	0.004	0.002
– Return comovement on data connectedness	0.003	0.005
ATT DiD coefficient of		
– App performance comovement on data connectedness	-0.023	-0.010
– Financial performance comovement on data connectedness	-0.003	-0.002
– Return comovement on data connectedness	-0.003	-0.004

B. Parameters			
Description	Source	Notation	Parameter Value
Data economy quarterly growth rate	Moment Matching	β	0.090
Post-ATT fraction of data linked to universal IDs	Moment Matching	ξ^{post}	0.700
Pre-ATT fraction of data linked to universal IDs	Normalization	ξ^{pre}	1.000
Quarterly data depreciation rate	External	μ	-0.075
Quarterly discount rate	External	ρ	0.030
Quarterly cash flow per unit of customer engagement	Data	ζ	18
Volatility of customer activities	Data	$\{\sigma_{\delta,i}\}_{i=1}^{N}$	In-text discussion
Data network matrix	Data	$\{\gamma_{ij}\}_{i,j=1}^{N}$	In-text discussion

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, includes 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 calibrated parameters.

Given these parameter values, 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 is run for 20 quarters which is our sample length in Section 3.

Using the simulated data, we replicate the regressions reported in Column (4) of Table 2, Table 3, and Table 4 and report the median estimates in the top panel of Table 6 alongside with the estimates from the aforementioned tables. These regressions target the impact of data connectedness on comovement in firms' operational, financial, and stock-market performances, before and after ATT, respectively. Note that there are six regression coefficients but only two parameters, β and ξ, that we adjust to match these coefficients. The comparison in Table 6 shows that our model captures the comovements in firms' operational and stock performances reasonably well.[42] In particular, the model successfully features a strong positive association between firms' performance

The SDKs adopted other methods of linking data from different firms. While such methods are imperfect substitutes for IDFA, they partially restored cross-app data linkage. The true disruption is less severe than the opt-out rate suggests.

[42]Market participants may not be informed about firms' data connectedness, so our valuations may not map perfectly to stock-market valuations. This partly explains the discrepancy between our model-implied and data moments.

comovements and their data-sharing linkages, and the ATT shock weakens this mechanism.[43]

4.4 Systemically important firms in data economy

The reduced-form findings in Section 3.1 (and reproduced in Section 4.3) are based on firms' direct connections through data sharing. Next, we apply our model to examine higher-order spillover effects: impact of a firm's data on peer firms is likely to transmit further to these firms' data-sharing counterparts, resulting in higher-degree externalities. In addition, as illustrated in Figure 6, data generates self-reinforcing growth, so a firm's data has a persistent effect over time on both itself and its connected peers by affecting the trajectory of data growth. Our valuation metrics comprehensively summarize the network externalities of higher orders and over multiple time horizons and thus allow us to identify systemically important firms in the data economy.

Aggregating the valuation of all firms, we obtain

$$\bar{V}_t = \zeta \mathbf{1}^\top \left(\hat{\rho}\mathbf{I} - \beta\Gamma\right)^{-1} \bar{\delta}_t + \sum_i v_{i,0}, \tag{17}$$

Firm i's contribution to the aggregate value of cash flows is given by

$$\zeta \{\mathbf{1}^\top \left(\hat{\rho}\mathbf{I} - \beta\Gamma\right)^{-1}\}_{.i} \delta_{i,t} + v_{i,0}, \tag{18}$$

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.[44]

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.,

$$\text{VKP} = \arg\max_i \ \zeta \{\mathbf{1}^\top \left(\hat{\rho}\mathbf{I} - \beta\Gamma\right)^{-1}\}_{.i} \delta_{i,t} + v_{i,0}. \tag{19}$$

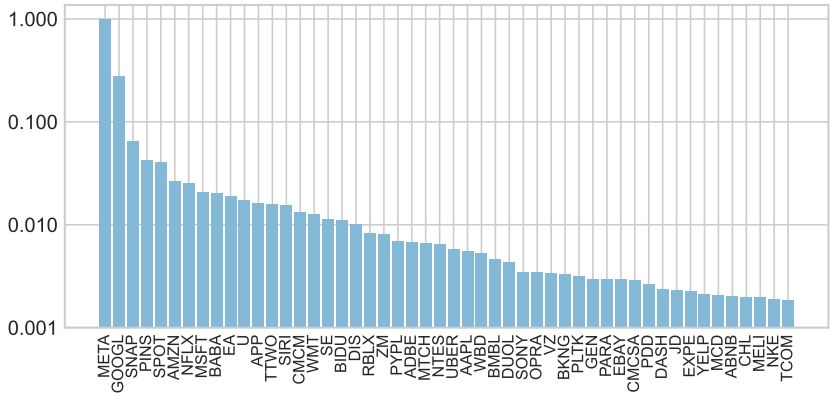
In Panel B of Figure 7, we report the valuation contribution from the top 50 firms ranked by their DAU and compare it against firms' DAU in Panel A. In both panels, we rank firms by DAU so the bar chart in Panel A exhibits a monotonically decreasing pattern, and we normalize firms' DAU and valuation contribution by that of the highest-ranked firm. Comparing the two panels reveals that firms' valuation contribution, which takes into account the data spillover effects of

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

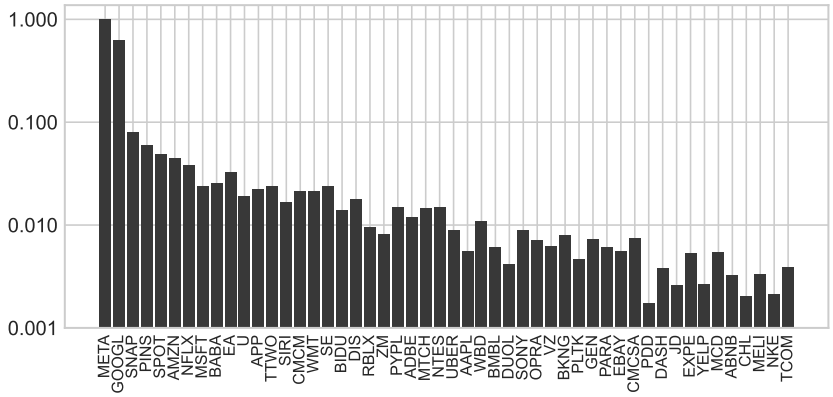
[44] Our focus on the scope of data usage across firms echoes Crouzet et al. (2024), who emphasize the scope of intangible capital usage across different divisions within firms and by competing imitators.

FIGURE 7: **Systemically Important Firms**

A. Top DAU firms



B. Valuation Contribution



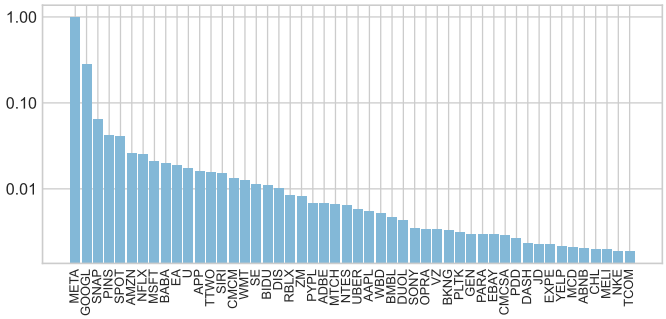
NOTE.—Figure 7 Panel A displays the top 50 firms ranked by their (log) average daily active users (DAU) within the sample. Firms are ordered by their average DAU from left to right. Panel B shows each firm's contribution to the aggregate valuation. Firm ranking is the same as in Panel A. The y-axis represents average DAU in Panel A and valuation contribution in Panel B, with the highest value normalized to 1.

all degrees of network propagation and across all time horizons, can differ significantly from firm size. Therefore, it is critical to account for the network structure of data flows.

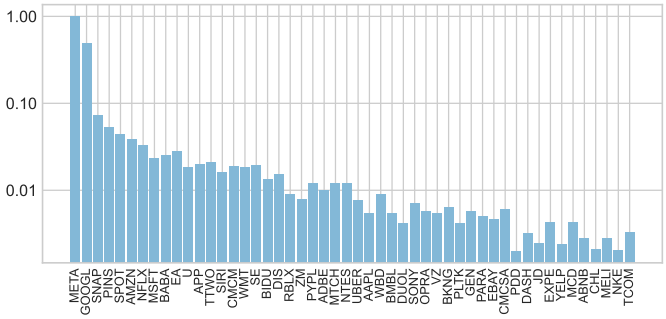
As shown in (16), a firm's contribution to the value of aggregate cash flows given by (18) can be decomposed into rounds of network propagation. At time t, the stock of firm i's raw data is $\delta_{i,t}$. By contributing to its own and other firms' information on customers, variation of $\delta_{i,t}$ permeates across the data-sharing network, generating direct ($k = 1$) and indirect ($k > 1$) spillover effects and, through the self-reinforcing data growth as illustrated in Figure 6, such impact persists into the future. In Figure 8, we truncate k at different values, denoted by K and report firms' valuation contribution. At $K = 0$, firm i's contribution to the value of aggregate cash flows is given by $\frac{\zeta}{\hat{\rho}}\delta_{i,t}$, which shuts down network propagation. In Panel B and C, we consider $K = 3$ and $K = \infty$, respectively. As K increases, valuation contributions converge to the equilibrium values.

FIGURE 8: **Valuation Contribution under Different Degrees of Network Propagation**

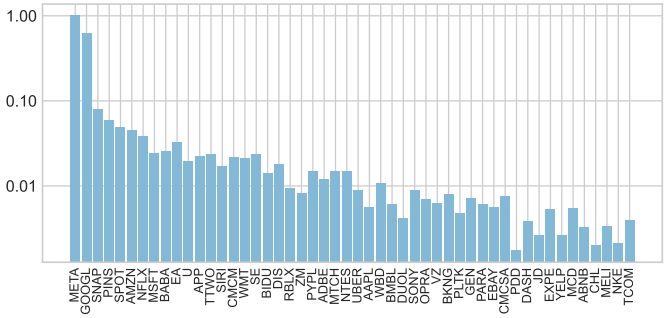
A. Valuation Contribution, $K = 0$



B. Valuation Contribution, $K = 3$



C. Valuation Contribution, $K = \infty$



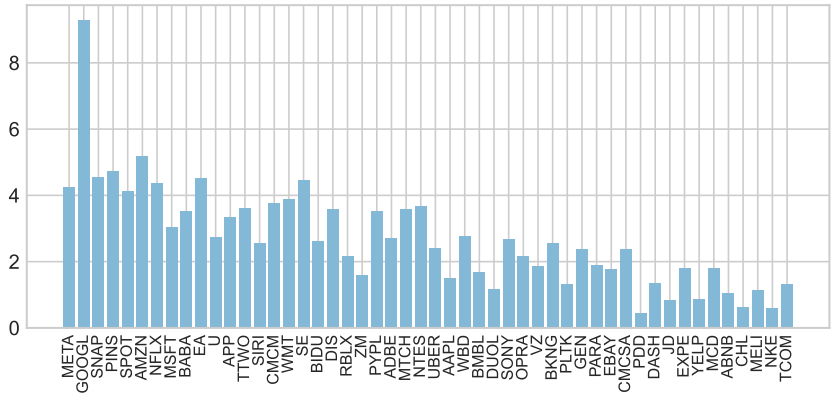
NOTE.—Figure 8 illustrates the distribution of valuation contribution at different levels of network propagation. The y-axis (in log scale) represents each firm's contribution to the total valuation of the network, with values scaled so that the maximum is normalized to 1. We display the top 50 firms 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.

A firm's contribution to the value of aggregate cash flows differs from its own valuation. In Figure 9, we report the ratio of valuation contribution to firm's own valuation:

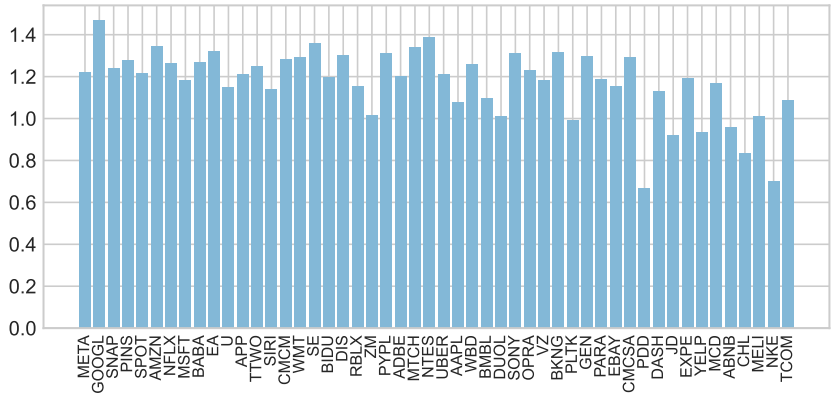
$$\frac{\zeta\{\mathbf{1}^\top \left(\hat{\rho}\mathbf{I} - \beta\Gamma\right)^{-1}\}_{.i}\delta_{i,t} + v_{i,0}}{\zeta \mathbf{e}_i^\top \left(\hat{\rho}\mathbf{I} - \beta\Gamma\right)^{-1}\bar{\delta}_t + v_{i,0}}. \tag{20}$$

FIGURE 9: **Firm Valuation vs. Firm Contribution to Aggregate Valuation**

A. Pre-ATT



B. Post-ATT



NOTE.—Figure 9 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 Apple Tracking Transparency (ATT) policy, and Panel B shows the corresponding distribution after ATT.

In Panel A, we compute the ratios under the pre-ATT value of ξ $(= 1)$, and in Panel B, we compute the post-ATT ratios under $\xi = 0.7$. In both panels, firms are sorted from left to right based on their DAU. Pre-ATT, contribution from Meta (formerly Facebook) to the value of aggregate cash flows in the data economy is more than 4.2x Meta's own valuation, and Alphabet has an even higher ratio above 9.3x. After the introduction of ATT that curtails the cross-firm data flows, Meta's and Alphabet's ratios declined to 1.2x and 1.5x respectively.

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 3. 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 Proposition 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 show that other firms' data, $\delta_{j,t}$, enters into firm i's choice of $x_{i,t}$ via firm i's information on its customers, $D_{i,t}$. Next, we show that $x_{i,t}$ is increasing in $D_{i,t}$, and herding behavior in firms' product-design decisions emerges due to data sharing. ATT weakens the behavior. Appendix E provides the proof.

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 ξ, i.e., $\frac{\partial^2 \mathbb{E}_t[dx_{i,t}]}{\partial x_{j,t} \partial \xi} > 0$.*

Consider an increase in other firms' current product design choice of $x_{j,t}$, which leads to an increase in their data stock, $\delta_{j,t}$, and an increase from t to $t + dt$ in firm i's information on its customers, $D_{i,t}$. Being more informed allows firm i to raise customer engagement via a higher $x_{i,t}$ without sacrificing monetization as much—that is, data alleviates the tension in product design between customer engagement and monetization as we have discussed in Section 4.2. Therefore, 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 in 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 suggest that waves of active data collection or monetization emerge in the data economy. Shocks are propagated across firms through data sharing, and as a result, other firms' marginal costs of data investment are interconnected. Positive shocks to one firm lead 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.[45] 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 ξ (the fraction of data linked to universal IDs), ATT weakens the cross-firm momentum in product-design decisions, i.e., $\frac{\partial \mathbb{E}_t[dx_{i,t}]}{\partial x_{j,t} \partial \xi} > 0$. This is an unintended consequence of ATT that has not been studied before. Beyond the impact on herding in product design, reducing ξ directly reduces firms' incentive to acquire data (i.e., decreases data marginal q) by weakening the data-driven firm growth, as ξ is part of the composite growth parameter β introduced in (14). As a result, ATT encourages firms to prioritize monetization. This prediction is in line with the findings in Kesler (2023). The next proposition summarizes this result.

Proposition 4 (ATT and Product Design) *Reducing ξ reduces the data marginal q, $\frac{\partial V_{i,t}}{\partial \delta_{i,t}}$, thereby causing all firms to prioritize monetization over customer engagement (i.e., $x_{i,t}$ decreases), and reduces $\frac{\partial \mathbb{E}_t[dx_{i,t}]}{\partial x_{j,t}}$, weakening the cross-firm momentum in product-design decisions.*

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 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 that support data security, customer service, and user reviews and feedback are associated with customer engagement. Examples of those functionality SDKs are provided in Subsection B.2. For each of the following SDK categories–payment, security, customer support, and review–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

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

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

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

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

	Monetization	User Engagement		
Feature:	Payment	Security	Support	Review
	(1)	(2)	(3)	(4)
L1.product feature (peers)	0.007***	0.003***	0.001***	0.001***
	(5.69)	(5.25)	(4.17)	(3.96)
ATT × L1.product feature (peers)	−0.011***	−0.002	−0.002***	−0.002***
	(−3.90)	(−1.20)	(−2.83)	(−3.44)
Industry#Quarter FE	Y	Y	Y	Y
Firm FE	Y	Y	Y	Y
Firm controls	Y	Y	Y	Y
Observations	18,596	18,596	18,596	18,596
R-sq	0.068	0.086	0.070	0.066

NOTE.—Table 7 presents evidence of cross-firm momentum in data accumulation investments. The outcome variable is the change in a firm's product design, measured by the number of unique functionality SDKs, as defined in Section 5.2. Column 1 reports the results on payment SDK, and Columns 2-4 report the results on SDKs that are likely to improve user engagement. Standard errors are clustered by firm. t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

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

$$\Delta X_{i,t} = \alpha + \beta_1 X_{-i,t-1} + \beta_2 \text{ATT} \times X_{-i,t-1} + \beta_3 X_{i,t-1} + \beta_4' \mathbf{C_{it}} + \iota_{st} + \varepsilon_{i,t}, \tag{21}$$

where $\mathbf{C_{it}}$ includes the same set of firm-level controls as in Equation (3). We also include industry-year fixed effects (ι_{st}). Standard errors are clustered at the firm level. The coefficient 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 Column 1 focusing on monetization SDKs and Columns 2-4 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 features. 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.

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	-0.007

NOTE.—Table 8 presents the regression coefficients from both the model and the data. The regression specification follows Equation (21). A positive coefficient indicates that herding behavior in firms' product 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 calibrated as in Table 6.

6 Conclusion

One firm's data can be used simultaneously by other firms and reveals information about other firms' customers. Such non-rival nature and externalities make data a uniquely productive asset. This paper uncovers a network of inter-firm data conduits, facilitated by data-analytics software. Firms collect data on their customers and share data 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.

Importantly, we develop a valuation framework for data assets that incorporates the data spillover effects of higher orders and across multiple time horizons. Based on our valuation metrics, we identify systemically important firms that hold critical positions in the data-sharing network and thus disproportionately influence the data economy. These findings highlight the need to consider the network structure of data flows 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., H. Tang, and B. Bian (2025). Breaking the Data Chain: The Ripple Effect of Data Sharing Restrictions on Financial Markets. *Available at SSRN 5334566*.

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, E. Oberfield, and V. Venkateswaran (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 (2023). The Effect of Privacy Regulation on the Data Industry: Empirical Evidence from GDPR. *RAND Journal of Economics 54*(4).

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.

Baker, S. R., B. Baugh, and M. Sammon (2023). Customer Churn and Intangible Capital. *Journal of Political Economy Macroeconomics 1*(3), 447–505.

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.

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., A. Bonatti, and T. Gan (2022). The Economics of Social Data. *The RAND Journal of Economics 53*(2), 263–296.

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.

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.

Buera, F. J. and E. Oberfield (2020). The Global Diffusion of Ideas. *Econometrica 88*(1), 83–114.

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.

Chen, D. (2024). The Market for Attention. Working paper.

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.

Comin, D. A., M. Dmitriev, and E. Rossi-Hansberg (2012, November). The Spatial Diffusion of Technology. Working Paper 18534, National Bureau of Economic Research.

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.

Corrado, C., C. Hulten, and D. Sichel (2005, January). Measuring Capital and Technology: An Expanded Framework.

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.

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.

Dong, D., A. Hu, Z. Li, and Z. Liu (2025). Information Acquisition and the Finance-Uncertainty Trap. Working paper, Federal Reserve Bank of San Francisco.

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.

Einav, L., P. J. Klenow, J. D. Levin, and R. Murciano-Goroff (2021, December). Customers and Retail Growth. Working Paper 29561, 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.

Goldberg, S. G., G. A. Johnson, and S. K. Shriver (2024). Regulating Privacy Online: An Economic Evaluation of the GDPR. *American Economic Journal: Economic Policy 16*(1), 325–358.

Gourio, F. and L. Rudanko (2014a). Can Intangible Capital Explain Cyclical Movements in the Labor Wedge? *American Economic Review 104*(5), 183–88.

Gourio, F. and L. Rudanko (2014b). 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.

Hall, B. H., J. Mairesse, and P. Mohnen (2010). Measuring the Returns to R&D. In B. H. Hall and N. Rosenberg (Eds.), *Handbook of the Economics of Innovation*, Volume 2, Chapter 24, pp. 1033–1082. Amsterdam: Elsevier.

Hayashi, F. (1982). Tobin's Marginal q and Average q: A Neoclassical Interpretation. *Econometrica 50*(1), 213–224.

He, B., L. I. Mostrom, and A. Sufi (2024). Investing in Customer Capital. Technical report, National Bureau of Economic Research.

Herskovic, B. (2018). Networks in Production: Asset Pricing Implications. *The Journal of Finance 73*(4), 1785–1818.

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.

Huo, Z., A. A. Levchenko, and N. Pandalai-Nayar (2025). International Comovement in the Global Production Network. *Review of Economic Studies 92*(1), 365–403.

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.

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.

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*.

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 (2025). Data Sales and Data Dilution. *Journal of Financial Economics 169*, 104053.

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.

Redding, S. J. and E. Rossi-Hansberg (2017). Quantitative Spatial Economics. *Annual Review of Economics 9*(Volume 9, 2017), 21–58.

Sims, C. A. (2003). Implications of Rational Inattention. *Journal of Monetary Economics 50*(3), 665–690. Swiss National Bank/Study Center Gerzensee Conference on Monetary Policy under Incomplete Information.

Varian, H. (2019). Artificial Intelligence, Economics, and Industrial Organization. In A. Agrawal, J. Gans, and A. Goldfarb (Eds.), *The Economics of Artificial Intelligence: An Agenda*, Chicago Scholarship Online. Chicago: University of Chicago Press.

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

Veldkamp, L. and J. Wolfers (2007). Aggregate Shocks or Aggregate Information? Costly Information and Business Cycle Comovement. *Journal of Monetary Economics 54*, 37–55.

Veldkamp, L. L. (2005). Slow Boom, Sudden Crash. *Journal of Economic Theory 124*(2), 230–257.

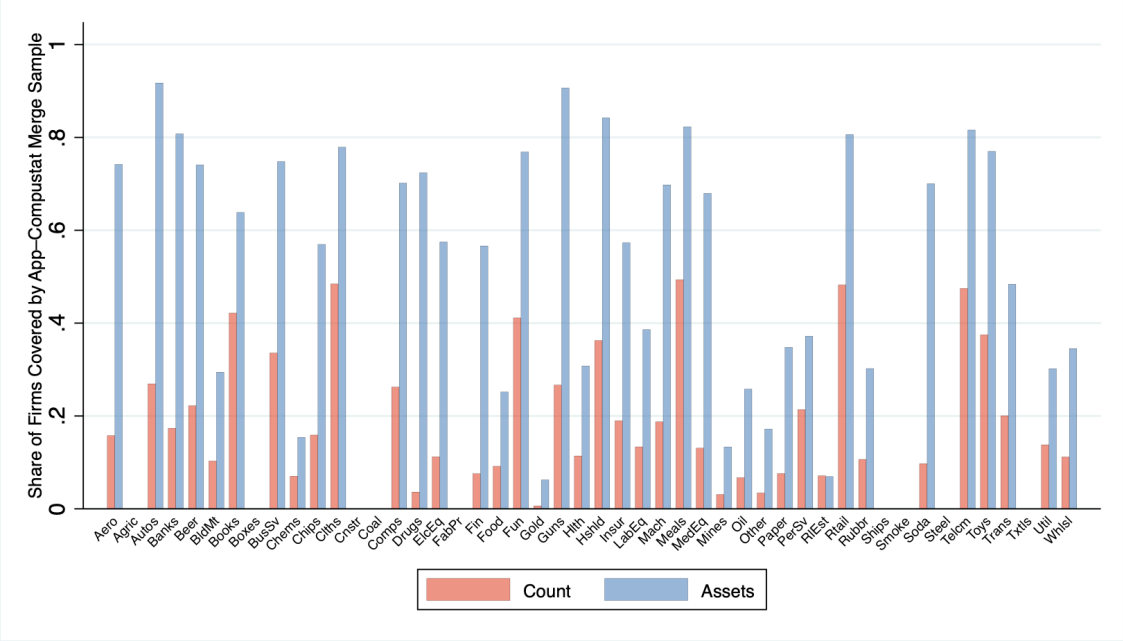
Veldkamp, L. L. (2006). Information Markets and the Comovement of Asset Prices. *The Review of Economic Studies 73*(3), 823–845.

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*.

Internet Appendix to
"Data as a Networked Asset"

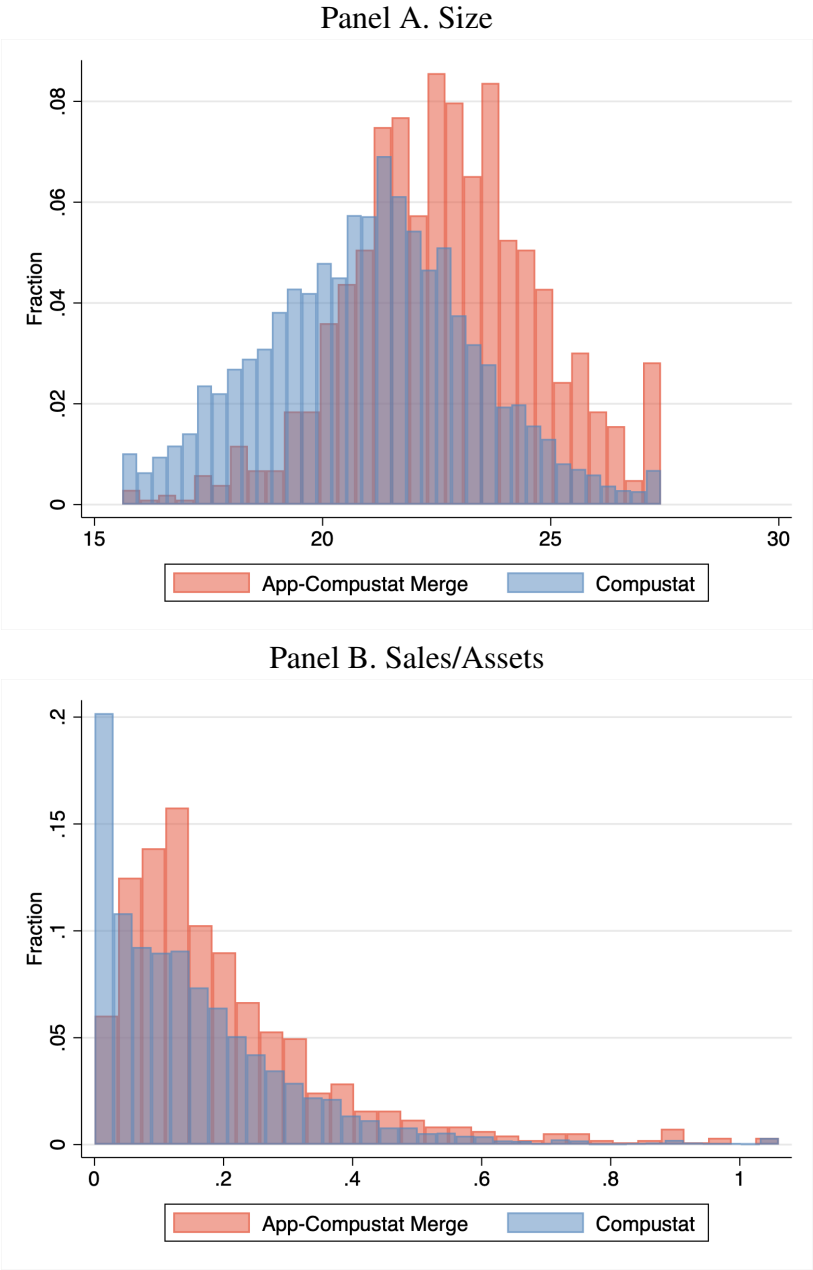
A Additional Figures and Tables

FIGURE A.1: **Firms in Our Sample vs. Compustat Universe by Industry**



NOTE.—Figure A.1 compares the industry distribution of firms in the data network ("App-Compustat Merge") with those in the broader Compustat universe ("Compustat"). Specifically, it shows the proportion of data-reliant firms within each Fama-French 48 industries.

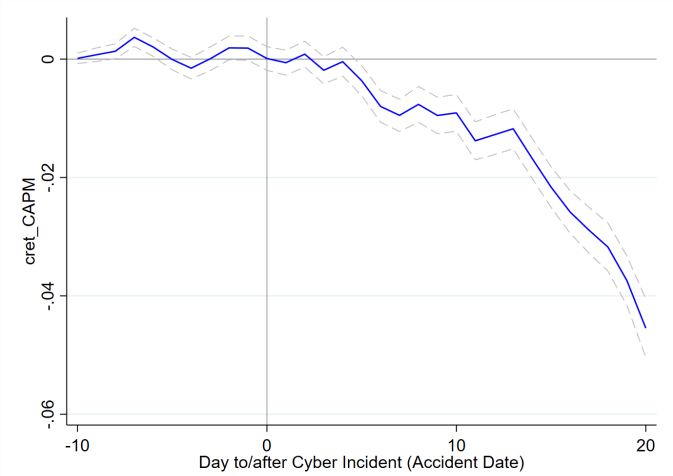
FIGURE A.2: **Firms in the Data Network vs. Compustat Universe: Size and Asset Turnover**



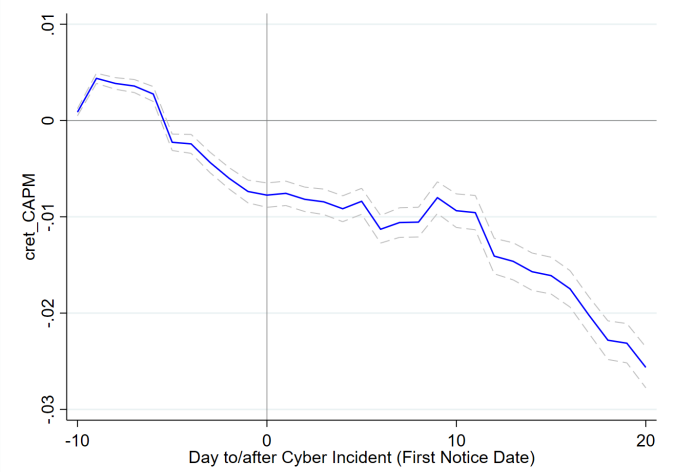
NOTE.—Figure A.2 compares firms in our sample ("App-Compustat Merge") with those in the broader Compustat universe ("Compustat"). We focus on firm size, proxied by log(assets) (Panel A) and the sales/assets ratio (Panel B).

FIGURE A.3: **Cyberattack Spillover Effects: Event Studies**

Panel A. Event Date: Incident Date



Panel B. Event Date: Notice Date



NOTE.— Figure A.3 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 75$^{\text{th}}$ percentile of the exposure distribution.

TABLE A.1: **Comovement in App Performance: Full Results**

	downloads		DAU	
	(1)	(2)	(3)	(4)
data connectedness	0.026***	0.024***	0.026***	0.023***
	(7.56)	(7.25)	(6.78)	(6.26)
ATT × data connectedness	-0.025***	-0.025***	-0.024***	-0.023***
	(-6.63)	(-6.74)	(-5.81)	(-5.75)
mobile user (0/1)		0.001		0.023***
		(0.79)		(7.44)
app category		0.008***		0.010***
		(4.85)		(5.64)
product horizontal		0.009***		0.010***
		(4.62)		(5.81)
product vertical		0.001		-0.002
		(0.36)		(-0.95)
supply chain (0/1)		-0.000		-0.001
		(-0.57)		(-1.60)
technology		-0.001		0.001
		(-1.42)		(0.56)
common analyst (0/1)		0.005***		0.005***
		(4.95)		(4.06)
geography		0.003		0.002
		(1.57)		(0.93)
ATT × mobile user (0/1)		0.010***		-0.004
		(4.06)		(-1.00)
ATT × app category		0.000		-0.003
		(0.00)		(-0.93)
ATT × product horizontal		-0.001		-0.005*
		(-0.29)		(-1.91)
ATT × product vertical		0.003		0.011*
		(0.72)		(1.88)
ATT × supply chain (0/1)		0.000		0.001
		(0.20)		(0.70)
ATT × technology		0.000		-0.002
		(0.38)		(-1.34)
ATT × common analyst (0/1)		-0.002		-0.001
		(-1.39)		(-0.81)
ATT × geography		0.002		0.000
		(0.77)		(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	0.068	0.113	0.114

NOTE.—Table A.1 shows the relationship between firm data connectedness and the comovement of app performance. Each observation represents a firm pair at a point in time. For each pair, the comovement of app performance is measured as the correlation between their quarterly log(downloads) and between their log(DAU), calculated separately for the periods before and after ATT (2021Q2). The data on app performance spans from 2014Q3 to 2023Q2. In even-numbered columns, we control for a comprehensive set of pairwise firm linkages, as well as interaction terms between these linkages and the ATT indicator, which equals one for periods after ATT. We include firm-by-time fixed effects (θ_{it} and ι_{jt}) and double-cluster standard errors by firm-i and -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 A.2: **Comovement in Firm Financial Performance: Full Results**

	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 j#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 A.2 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 pairwise 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 A.3: **Comovement in Stock Returns: Full Results**

	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***
	(6.52)	(3.64)	(6.71)	(4.35)	(6.25)	(3.76)
ATT × data connectedness	-0.002***	-0.002***	-0.003***	-0.003***	-0.002**	-0.002***
	(-3.00)	(-3.05)	(-3.86)	(-3.95)	(-2.53)	(-2.67)
mobile user (0/1)		0.001***		0.002***		0.001*
		(4.51)		(4.22)		(1.77)
app category		0.008***		0.009***		0.004***
		(8.60)		(7.89)		(5.25)
product horizontal		0.022***		0.030***		0.017***
		(22.26)		(21.54)		(19.58)
product vertical		0.002*		0.001		0.007***
		(1.70)		(0.96)		(4.26)
supply chain (0/1)		0.000		0.000		0.000
		(0.60)		(0.90)		(1.31)
technology		0.001***		0.002***		0.001
		(3.53)		(3.98)		(1.45)
common analyst (0/1)		0.009***		0.012***		0.009***
		(17.03)		(16.69)		(17.28)
geography		0.004***		0.004***		0.003***
		(4.50)		(3.90)		(3.66)
ATT × mobile user (0/1)		-0.001***		-0.002***		-0.000
		(-4.61)		(-4.78)		(-1.03)
ATT × app category		-0.001		-0.002		-0.001
		(-0.67)		(-1.50)		(-1.07)
ATT × product horizontal		0.001		-0.001		-0.001
		(0.55)		(-0.86)		(-1.29)
ATT × product vertical		0.001		0.003*		-0.002
		(0.68)		(1.69)		(-1.25)
ATT × supply chain (0/1)		0.000		0.000		0.000
		(0.36)		(1.00)		(1.10)
ATT × technology		0.000		-0.000		0.000
		(0.60)		(-0.09)		(0.26)
ATT × common analyst (0/1)		0.001		0.002*		0.002***
		(1.29)		(1.87)		(2.69)
ATT × geography		-0.002		-0.002		0.000
		(-1.45)		(-1.15)		(0.26)
Firm i#Time FE	Y	Y	Y	Y	Y	Y
Firm j#Time FE	Y	Y	Y	Y	Y	Y
Observations	5,748,100	5,748,100	5,645,298	5,645,298	4,993,880	4,993,880
R-sq	0.442	0.449	0.097	0.110	0.022	0.028

NOTE.—Table A.3 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. t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

TABLE A.4: **Performance Comovement Test with Placebo Connectedness**

	downloads		sales/assets		return-CAPM	
	(1)	(2)	(3)	(4)	(5)	(6)
media SDK connectedness	0.003**		−0.000		0.001	
	(2.03)		(−0.12)		(0.54)	
ATT × media SDK connectedness	−0.002		0.000		0.003	
	(−0.83)		(0.61)		(1.50)	
message SDK connectedness		0.001		−0.001***		0.000
		(0.94)		(−6.07)		(−0.53)
ATT × message SDK connectedness		−0.000		0.002**		−0.001
		(−0.03)		(2.51)		(−0.53)
Firm i#ATT FE	Y	Y	Y	Y	Y	Y
Firm j#ATT FE	Y	Y	Y	Y	Y	Y
Other connections	Y	Y	Y	Y	Y	Y
Other connections#ATT	Y	Y	Y	Y	Y	Y
Observations	1,401,082	1,401,082	1,379,592	1,379,592	343,202	950,718
R-sq	0.067	0.067	0.005	0.005	0.100	0.137

NOTE.— Table A.4 shows the relationship between placebo connectedness and performance comovement. We construct placebo connectedness following the definition of Equation (1), replacing the top 50 data SDKs with the top 20 media SDKs (in odd columns) and message SDKs (in even columns). The outcome variables are comovement based on downloads (Columns 1–2), sales-to-assets (Columns 3–4), and CAPM-adjusted returns (Columns 5–6), respectively. In all 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. t-statistics are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

TABLE A.5: **Cyberattack Spillover Effects on App Performances: Full Results**

	log(downloads)		log(DAU)	
	(1)	(2)	(3)	(4)
cyber event × high exposure	−0.116***	−0.107***	−0.148***	−0.137***
	(−6.60)	(−5.31)	(−7.28)	(−5.75)
cyber event × mobile user (0/1)		−0.003		−0.000
		(−0.19)		(−0.01)
cyber event × app category		−0.020		−0.032*
		(−1.25)		(−1.82)
cyber event × product horizontal		−0.003		−0.005
		(−0.20)		(−0.26)
cyber event × product vertical		0.026		0.029
		(0.70)		(0.63)
cyber event × technology		−0.025*		−0.031*
		(−1.85)		(−1.96)
cyber event × supply chain (0/1)		−0.006		−0.010
		(−0.70)		(−0.86)
cyber event × common analyst (0/1)		0.000		0.006
		(0.01)		(0.34)
cyber event × geography		−0.001		0.008
		(−0.04)		(0.27)
Firm controls	Y	Y	Y	Y
Firm#Event FE	Y	Y	Y	Y
Event-specific relative quarter FE	Y	Y	Y	Y
Observations	81,373	81,373	81,373	81,373
R-sq	0.955	0.955	0.957	0.957

NOTE.—Table A.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 that result 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 75$^{\text{th}}$ 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 A.6: **Cyberattack Spillover Effects on Firms' Performances: Placebo Connectedness**

	log(downloads)		log(DAU)	
	(1)	(2)	(3)	(4)
cyber event × high exposure (media)	−0.003		−0.003	
	(−1.40)		(−1.24)	
cyber event × high exposure (msg)		−0.004		−0.002
		(−1.20)		(−0.63)
Firm controls	Y	Y	Y	Y
Firm#Event FE	Y	Y	Y	Y
Event-specific relative quarter FE	Y	Y	Y	Y
Observations	194,427	194,427	194,427	194,427
R-sq	0.958	0.958	0.962	0.962

NOTE.—Table A.6 reports the cross-firm spillover effects of major cyber events, using placebo connectedness to define firms exposed to these events. Placebo connectedness is constructed following the definition in Equation (1), substituting the top 50 data SDKs with the top 20 media SDKs (in odd-numbered columns) and message SDKs (in even-numbered columns). The specification follows that of Columns 2 and 4 in Table 5. All regressions control for firm characteristics, 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.

B Additional Information on Variable Construction

B.1 Additional Data Sources

Other App Metrics. We obtain quarterly app category and user overlap data from Apptopia. To construct the overlap in app category between firm pairs ("*app category*"), we create a binary vector recording a firm's presence in 27 iOS app categories and compute two firms' cosine similarity.

Mobile user overlap is a binary variable, defined as 1 if any app from firm i shares users with any app from firm j in the Apptopia data, and 0 otherwise ("*mobile user*"). Apptopia derives this metric from 1.5 billion user reviews, each linked to a unique store user ID, which allows identification of audience overlap when a user reviews multiple apps.

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). Horizontal industry linkages ("*product horizontal*") are constructed using pairwise firm similarity scores derived from text analysis of 10-K product descriptions. Vertical industry linkages ("*product vertical*") are based on product vocabularies from the Bureau of Economic Analysis (BEA) Input-Output tables, combined with firms' 10-K product descriptions. Both measures are available at an annual frequency.

Factset Revere. Supply chain relationships are constructed from FactSet Revere Supply Chain Relationships datasets. This data is collected by Factset based on companies' relationship information from primary public sources such as SEC 10-K annual filings, investor presentations, and press releases. For each firm, we create an indicator variable that equals to one ("*supply chain*") if the firm pair has customer-supplier relationships. We also construct a measure of geographical overlap using the FactSet Revere Geographic Revenue Exposure datasets, which provide revenue breakdowns by geography and business segment. The data are organized into a four-level geographic classification (Super-Region, Region, Area, Country). Our geographic overlap measure ("*geography*") is based on the third level, Area, which includes 29 distinct regions. We compute cosine similarity between firm pairs using their revenue distribution vectors across these areas. Both 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 follow the methodology of Jaffe (1986) and Bloom, Schankerman, and Van Reenen (2013) to compute technology proximity. Specifically, we first calculate the share of a firm's patents in each technology class and then compute the cosine similarity between firm pairs ("*technology*"). This measure is available at an annual frequency.

Cyberattack events. We obtain information on cyberattack events 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. We identify 22 major cyberattack events that involve at least 10 million exposed records and list them in the table at the end of this section.

Standard financial datasets. Firms that share sell-side analysts may exhibit performance co-movement (Ali and Hirshleifer, 2020). Using analyst coverage data from I/B/E/S, we construct a binary variable equal to one if a firm pair shares at least one common analyst, and zero otherwise ("*common analyst*").

Quarterly data on firm accounting variables is from Compustat and data on stock prices and market capitalization from CRSP. Data on asset pricing factors is from Ken French's data library.

TABLE B.1: Major Cyberattack 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.

B.2 Examples of Product-Design SDKs

TABLE B.2: Description of Example Product-Design (Functionality) SDKs

SDK Name	Category	# Installation	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.

TABLE C.1: **Summary Statistics: Additional Variables**

Panel A. Pairwise variables

	mean	sd	p10	p25	p50	p75	p90	count
Pairwise connections								
data connectedness	0.171	0.19	0.00	0.00	0.11	0.29	0.45	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

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)	4.407	1.42	2.96	4.45	4.91	5.15	5.32	20,344
L1.security SDK (peers)	4.442	1.43	2.97	4.49	4.95	5.20	5.36	20,344
L1.customer support SDK (peers)	3.865	1.30	2.17	3.80	4.28	4.59	4.80	20,344
L1.review & feedback SDK (peers)	4.144	1.36	2.62	4.14	4.60	4.88	5.05	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	22.789	2.02	20.22	21.37	22.75	24.13	25.51	18,596
L1.long-term debt/assets	0.269	0.23	0.01	0.08	0.23	0.39	0.57	18,596
L1.tangibles/assets	0.206	0.22	0.01	0.04	0.12	0.31	0.57	18,596
L1.cash/assets	0.182	0.18	0.02	0.05	0.11	0.25	0.47	18,596
L1.R&D/assets	0.000	0.00	0.00	0.00	0.00	0.00	0.00	18,596
L1.investments/assets	0.021	0.03	0.00	0.00	0.01	0.03	0.05	18,596

NOTE.—Table C.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 September 2014 to June 2023. Panel B includes variables constructed at the firm level, all at quarterly frequency.

C Additional Summary Statistics

Other statistics. Table C.1 provides summary statistics for all the variables in our sample, which includes 1,031 firms over the period from 2014Q3 to 2023Q2. Panel A lists all firm linkages. The app-category similarity and similarity in geographical distribution of business segments 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.

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 has a long term debt ratio of 26.2%, a tangibility of 20.6%, and a cash-to-asset ratio of 18.2%.

D Performance Comovement: Robustness Checks

We demonstrate that the results on performance comovement are robust to alternative measures of data connectedness and regression specifications. For each robustness test, we plot in Figure D.1 the estimated coefficients of data connectedness (left panel) and the coefficients on the interaction between data connectedness and ATT indicator variable (right panel) from Equation (2), focusing on correlations based on three metrics of firms' performances—app downloads, sales-to-assets, and CAPM-adjusted returns—shown from top to bottom panels.

Alternative measures of data connectedness. We consider six variants of the data connectedness measure. First, we compute data connectedness using only the top 20 data-related SDKs, ranked by cumulative installations. In the main text, we consider the top 50 data-related SDKs.

Second, we incorporate the popularity of each SDK into the construction of the data connectedness measure by extending S_{jt} to $\widetilde{S_{jt}} = \Omega_t \times S_{jt}$, where Ω_t is a diagonal matrix whose diagonal element ω_{it} represents the market share of SDK i at time t:

$$\Omega_t = \begin{pmatrix} \omega_{1t} & 0 & \cdots & 0 \\ 0 & \omega_{2t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_{Kt} \end{pmatrix}.$$

The resulting *popularity-weighted* ("PW") data connectedness is defined as

$$\rho_{ijt}^{\text{data, PW}} = \frac{\widetilde{S_{it}}' \widetilde{S_{jt}}}{|\widetilde{S_{it}}| \cdot |\widetilde{S_{jt}}|}.$$

Third, we incorporate heterogeneity in the data collection intensity of each app by modifying the elements of the SDK usage vector S_{ikt}. Specifically, we redefine $s_{iakt} = m_{iat} \times d_{iakt}$ (in the main text) as $s_{iakt} = m_{iat} \times d_{iakt} \times n_{iat}$, where n_{iat} is the number of unique data types collected by app a based on its privacy labels (following the definition in Bian et al. (2021)).[1]

The next three variants address the concern that our results depend on how to time-average ρ_{ijt}^{data}. In the main text, we consider the pre-ATT average values of ρ_{ijt}^{data}. If the data network is relatively stable, we would expect the results to hold even when averaging data connectedness in alternative time periods rather than in the pre-ATT period. For the fourth variant of data connectedness measure, we use the average quarterly data connectedness in 2020 (the year before ATT was introduced). In the fifth variant, we use the average in the post-ATT period. Lastly, we compute averages for the pre-ATT and post-ATT periods separately so the left-side outcome variable is matched in time subscript ($\in \{$*pre-ATT*, *post-ATT*$\}$) with the right-side data connectedness.

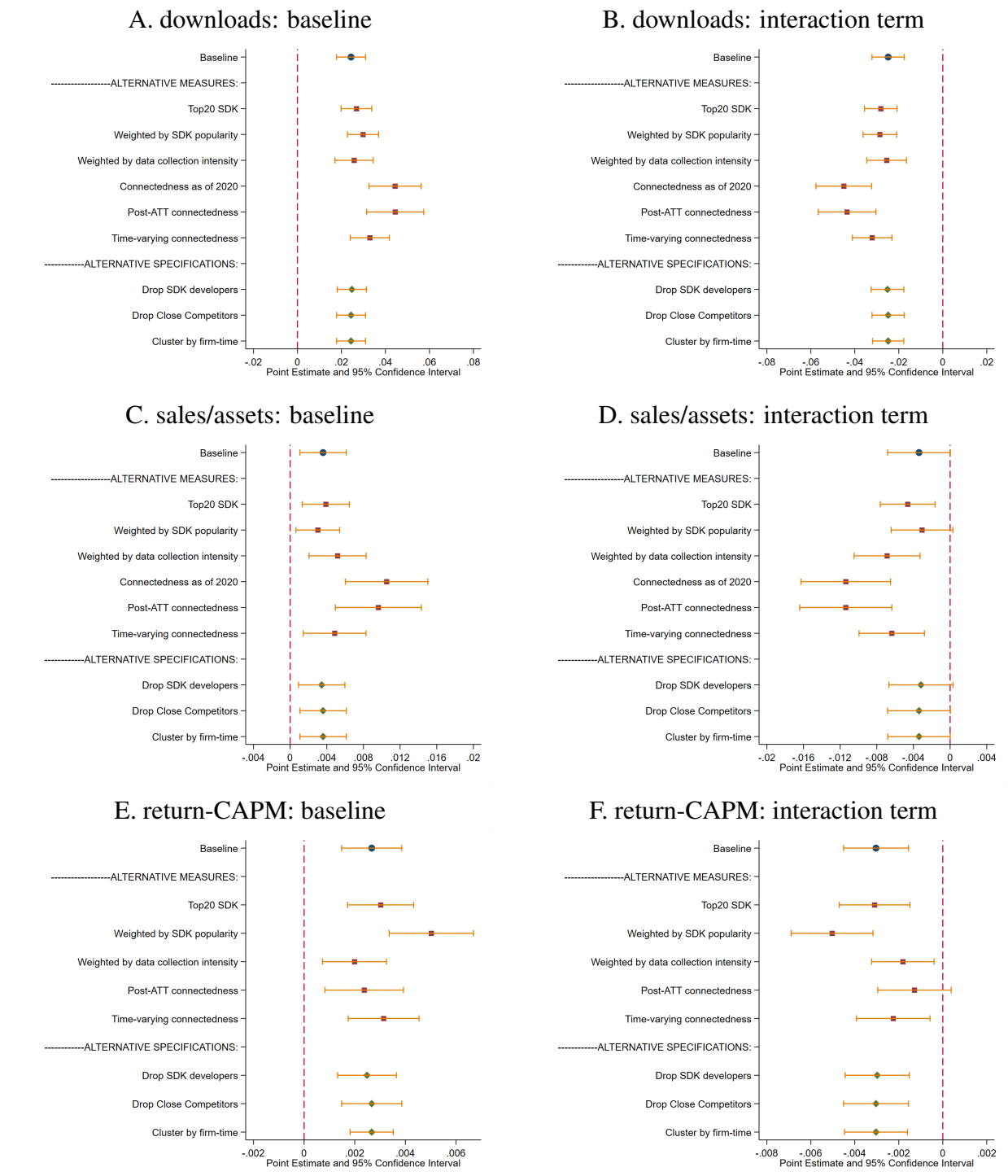
Alternative specifications. We consider three alternative regression specifications: (1) excluding eight firms that are data-related SDK providers—AppLovin, Twitter, Google, Meta, Yandex, Adobe, Unity, and Amazon; (2) dropping close competitors, defined as pairwise connection in

[1] An example of Facebook's privacy label is available at `https://apps.apple.com/us/app/facebook/id284882215`.

product space (Hoberg–Phillips horizontal product-space overlap) above the sample median; and (3) clustering standard errors at both the firm-i-by-time and firm-j-by-time levels. Note that the second specification is motivated by the potential concern that data-sharing relationships between direct competitors may be confounded by strategic considerations or contractual limitations.

Robustness analysis results. In summary, we consider a total of 9 variant regressions, including 6 data connectedness measures in the baseline regression specification and 3 regression specifications using the baseline data connectedness measure, so, with 3 outcome variables representing firms' app, financial, and stock-market performances, respectively, we have a total of 27 sets of robustness results. The estimates of coefficient of data connectedness (interaction term with ATT) are almost always positive (negative) and statistically significant at the 5% level. These findings suggest that our results are not driven by any factors above that motivate our robustness analysis.

FIGURE D.1: **Robustness Checks**



NOTE.—Figure D.1 presents the estimated baseline coefficients (left panel) and the coefficients on the interaction term (right panel) from Equation (2) across various alternative measures and specifications. From top to bottom, the correlation is based on downloads, sales-to-assets, and CAPM-adjusted returns, respectively.

E Model Derivation and Proofs

E.1 Proof of Proposition 1

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

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

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

$$-g_x(x_{i,t}, \lambda_{i,t}) = V^i_{\delta_{i,t}}\theta\,. \tag{E.2}$$

E.2 Proof of Proposition 2

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

$$V\left(\delta_{i,t}, \{\delta_{j,t}\}\right) = v_{i,0} + v_i^\top\bar{\delta}_t = v_{i,0} + \sum_{j=1}^{N} v_{i,j}\delta_{j,t}, \tag{E.3}$$

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 dependence 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:

$$\rho V\left(\delta_{i,t}, \{\delta_{j,t}\}\right)dt = \max_{x_{i,t}}\left(g\left(x_{i,t}, \lambda_{i,t}\right) + \zeta\delta_{i,t}\right)dt + v_{i,i}\left(\theta(x_{i,t} + \frac{\alpha\kappa}{\sigma_\eta}D_{i,t}) + \mu_\delta\delta_{i,t}\right)dt + v_{i,j}\sum_{j\neq i}\left[\theta(x_{j,t} + \frac{\alpha\kappa}{\sigma_\eta}D_{j,t}) + \mu_\delta\delta_{j,t}\right]dt\,. \tag{E.4}$$

FOC (12) becomes

$$\frac{\zeta\phi_1}{(\phi_0 D_{i,t} - \phi_1 x_{i,t})} = \theta v_{i,i}, \tag{E.5}$$

where $\phi_0 = \phi_\lambda\kappa/\sigma_\eta$. If $v_{i,i} > 0$, this ensures that $\phi_0 D_{i,t} - \phi_1 x_{i,t} > 0$ and also gives

$$x_{i,t} = \frac{\phi_0}{\phi_1}D_{i,t} - \frac{\zeta}{\theta v_{i,i}}. \tag{E.6}$$

That is

$$x_{i,t} = \frac{\phi_0}{\phi_1} D_{i,t} - \frac{\zeta}{\theta v_{i,i}} = \frac{\phi_0}{\phi_1} \xi (\sum_{j=1}^{N} \gamma_{ij} \delta_{j,t}) - \frac{\zeta}{\theta v_{i,i}}. \tag{E.7}$$

The cash flow from user activities is given by

$$g_{i,t} = \zeta \log(\frac{\zeta \phi_1}{\theta v_{i,i}}). \tag{E.8}$$

Substitute FOC into HJB

$$\begin{aligned} \rho V_{i,t} &= \zeta \delta_{i,t} + \zeta \log(\frac{\zeta \phi_1}{\theta v_{i,i}}) + \sum_{j=1}^{N} v_{i,j} \left[\theta x_{j,t} + \frac{\alpha \kappa}{\sigma_\eta} \theta D_{j,t} + \mu_\delta \delta_{j,t} \right] \\ &= \zeta \delta_{i,t} + \zeta \log(\frac{\zeta \phi_1}{\theta v_{i,i}}) + \sum_{j=1}^{N} v_{i,j} \left(\theta \frac{\phi_0}{\phi_1} D_{j,t} - \frac{\zeta}{v_{j,j}} + \frac{\alpha \kappa}{\sigma_\eta} \theta D_{j,t} + \mu_\delta \delta_{j,t} \right), \end{aligned} \tag{E.9}$$

where

$$D_{j,t} = \left(\sum_{k=1}^{N} \gamma_{jk} \delta_{k,t} \xi \right). \tag{E.10}$$

After substitution and simplification:

$$\rho V_{i,t} = \left(\zeta \delta_{i,t} + \sum_{j=1}^{N} v_{i,j} \left[\theta (\frac{\phi_0}{\phi_1} + \frac{\alpha \kappa}{\sigma_\eta}) \left(\sum_{k=1}^{N} \gamma_{jk} \delta_{k,t} \xi \right) + \mu_\delta \delta_{j,t} \right] \right) + A_i, \tag{E.11}$$

where

$$A_i = \left(\zeta \log(\frac{\zeta \phi_1}{\theta v_{i,i}}) + \sum_{j=1}^{N} v_{i,j} \left[-\frac{\zeta}{v_{j,j}} \right] \right). \tag{E.12}$$

This gives the constant term in the valuation

$$v_{i,0} = \frac{A_i}{\rho}. \tag{E.13}$$

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 (E.11), the coefficients on $\delta_{i,t}$ should be the same on both sides, therefore we have for $v_{i,i}$:

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

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

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

Define:

$$\beta = \theta(\frac{\phi_0}{\phi_1} + \frac{\alpha\kappa}{\sigma_\eta})\xi.$$

The above equations can be written as

$$(\rho - \mu_\delta) v_{i,i} = \zeta + \beta(\Gamma^\top v_i)_i \tag{E.14}$$

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:

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

Solving for the valuation vector v_i, we obtain

$$v_i = (I\hat{\rho} - \beta\Gamma^\top)^{-1} \zeta \mathbf{e}_i. \tag{E.15}$$

Therefore, the valuation vector is given by

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

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

$$V_{i,t} = v_{i,0} + \frac{\zeta}{\hat{\rho}}\mathbf{e}_i^\top\left(I + \frac{\beta}{\hat{\rho}}\Gamma + (\frac{\beta}{\hat{\rho}})^2\Gamma^2 + (\frac{\beta}{\hat{\rho}})^3\Gamma^3 + ...\right)\bar{\delta}_t \tag{E.16}$$

$$= v_{i,0} + \eta\delta_{i,t} + \eta\mathbf{e}_i^\top\frac{\beta}{\hat{\rho}}(I + \frac{\beta}{\hat{\rho}}\Gamma + (\frac{\beta}{\hat{\rho}})^2\Gamma^2 + ...)\Gamma\bar{\delta}_t \tag{E.17}$$

$$= v_{i,0} + \eta\delta_{i,t} + \eta\mathbf{e}_i^\top\frac{\beta}{\hat{\rho}}(I - \frac{\beta}{\hat{\rho}}\Gamma)^{-1}\mathbf{D}_t, \tag{E.18}$$

where the constant η is defined as

$$\eta = \frac{\zeta}{\hat{\rho}}. \tag{E.19}$$

From the valuation of firm i given by

$$V_{i,t} = v_{i,0} + \eta\delta_{i,t} + \eta\frac{\beta}{\hat{\rho}}\mathbf{e}_i^\top\sum_{k=0}^{\infty}\left(\frac{\beta}{\hat{\rho}}\Gamma\right)^k(\mathbf{D}_t), \tag{E.20}$$

we derive the network-augmented Gordon growth formula:

$$V_{i,t} = v_{i,0} + \eta\delta_{i,t} + \eta\frac{\beta}{\hat{\rho}}\mathbf{e}_i^\top\left(\mathbf{I} + \frac{\beta}{\hat{\rho}}\Gamma + (\frac{\beta}{\hat{\rho}})^2\Gamma^2 + ...\right)\Gamma\bar{\delta}_t \tag{E.21}$$

$$= v_{i,0} + \eta\mathbf{e}_i^\top\mathbf{I}\bar{\delta}_t + \eta\frac{\beta}{\hat{\rho}}\mathbf{e}_i^\top\left(\Gamma + \frac{\beta}{\hat{\rho}}\Gamma^2 + (\frac{\beta}{\hat{\rho}})^2\Gamma^3 + ...\right)\bar{\delta}_t \tag{E.22}$$

$$= v_{i,0} + \eta\mathbf{e}_i^\top\left(\mathbf{I} + \frac{\beta}{\hat{\rho}}\Gamma + (\frac{\beta}{\hat{\rho}})^2\Gamma^2 + (\frac{\beta}{\hat{\rho}})^3\Gamma^3 + ...\right)\bar{\delta}_t \tag{E.23}$$

$$= v_{i,0} + \zeta\mathbf{e}_i^\top(\hat{\rho}\mathbf{I} - \beta\Gamma)^{-1}\bar{\delta}_t. \tag{E.24}$$

E.3 Proof of Corollary 2

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

$$dR_{i,t} = v_i^\top\frac{d\bar{\delta}_t}{V_{i,t}} = \mathbf{e}_i^\top(I - \frac{\beta}{\hat{\rho}}\Gamma)^{-1}\frac{\zeta}{\hat{\rho}}\frac{d\bar{\delta}_t}{V_{i,t}} = \frac{\zeta}{\hat{\rho}}\sum_{n=1}^{N}\mathbf{e}_i^\top\left(I - \frac{\beta}{\hat{\rho}}\Gamma\right)^{-1}\mathbf{e}_n\frac{d\delta_{n,t}}{V_{i,t}}. \tag{E.25}$$

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

$$\rho_{ij} = \text{corr}(dR_{i,t}, dR_{j,t}) = \frac{\text{cov}(dR_{i,t}, dR_{j,t})}{\sqrt{\text{var}(dR_{i,t})\text{var}(dR_{j,t})}}. \tag{E.26}$$

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,

$$\begin{aligned} \text{corr}(dR_{i,t}, dR_{j,t}) &= \frac{\text{cov}\left(\sum_{n=1}^N \mathbf{e}_i^\top \mathbf{M}\mathbf{e}_n \frac{d\delta_{n,t}}{V_{i,t}}, \sum_{n=1}^N \mathbf{e}_j^\top \mathbf{M}\mathbf{e}_n \frac{d\delta_{n,t}}{V_{j,t}}\right)}{\sqrt{\text{var}\left(\sum_{n=1}^N \mathbf{e}_i^\top \mathbf{M}\mathbf{e}_n \frac{d\delta_{n,t}}{V_{i,t}}\right)\text{var}\left(\sum_{n=1}^N \mathbf{e}_j^\top \mathbf{M}\mathbf{e}_n \frac{d\delta_{n,t}}{V_{j,t}}\right)}} \qquad \text{(E.27)} \\ &= \frac{\text{cov}\left(\sum_{n=1}^N \mathbf{e}_i^\top \mathbf{M}\mathbf{e}_n \sigma_{n,\delta}\delta_{n,t}dz_{n,t}, \sum_{n=1}^N \mathbf{e}_j^\top \mathbf{M}\mathbf{e}_n \sigma_{n,\delta}\delta_{n,t}dz_{n,t}\right)}{\sqrt{\text{var}\left(\sum_{n=1}^N \mathbf{e}_i^\top \mathbf{M}\mathbf{e}_n \sigma_{n,\delta}\delta_{n,t}dz_{n,t}\right)\text{var}\left(\sum_{n=1}^N \mathbf{e}_j^\top \mathbf{M}\mathbf{e}_n \sigma_{n,\delta}\delta_{n,t}dz_{n,t}\right)}}. \qquad \text{(E.28)} \end{aligned}$$

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 (E.28) 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 (E.28) 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

$$\rho_{ij} = \frac{\mathbf{M}_i^\top \Sigma_z \mathbf{M}_j}{\|\mathbf{M}_i\| \|\mathbf{M}_j\|} \tag{E.29}$$

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 $\mathbf{M} = (I - \frac{\beta}{\hat{\rho}}\Gamma)^{-1}$:

$$\frac{\partial \mathbf{M}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} \mathbf{M} E_{ij} \mathbf{M}, \tag{E.30}$$

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

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

$$\frac{\partial M_{k,l}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}}(\mathbf{M}E_{ij}\mathbf{M})_{k,l} = \frac{\beta}{\hat{\rho}}M_{k,i}M_{j,l}. \tag{E.31}$$

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 (E.31) to calculate the derivatives:

$$\frac{\partial M_{i,n}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} M_{i,i} M_{j,n}, \tag{E.32}$$

$$\frac{\partial M_{j,n}}{\partial \Gamma_{i,j}} = \frac{\beta}{\hat{\rho}} M_{j,i} M_{j,n}. \tag{E.33}$$

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 an 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, $\frac{\partial^2 \rho_{ij}}{\partial \Gamma_{i,j} \partial \beta} > 0$.

E.4 Proof of Proposition 3

For firm i, substituting the functional form of its valuation (E.3) into the HJB equation, then taking the FOC with respect to $x_{i,t}$, we get

$$x_{i,t} = \frac{\phi_0}{\phi_1} D_{i,t} - \frac{\zeta}{\theta v_{i,i}}, \tag{E.34}$$

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 (E.15). Recall that the law of motion of data capital is given by

$$d\delta_{i,t} = \left(\theta x_{i,t} + \alpha\theta \frac{\kappa}{\sigma_\eta} D_{i,t} + \mu_\delta \delta_{i,t} \right) dt + \sigma_{i,\delta} \delta_{i,t} dz_{i,t}. \tag{E.35}$$

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

$$dx_{i,t} = \frac{\phi_0}{\phi_1} \left[\sum_{j=1}^{N} \gamma_{ij} \xi \left((\theta x_{j,t} + \alpha\theta \frac{\kappa}{\sigma_\eta} D_{j,t} + \mu_\delta \delta_{j,t}) dt + \sigma_{j,\delta} \delta_{j,t} dz_{j,t} \right) \right]. \tag{E.36}$$

Therefore we obtain $\frac{\partial^2 \mathbb{E}[dx_{i,t}]}{\partial x_{j,t} \, \partial \xi} = \frac{\phi_0}{\phi_1} \gamma_{ij} \, \theta > 0$.

E.5 Proof of Proposition 4

Recall that in E.2 we show the firm's valuation takes the following form

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

Hence

$$V_{\delta_{i,t}} := \frac{\partial V_{i,t}}{\partial \delta_{i,t}} = \zeta \, \left[(\hat{\rho} I - \beta \, \Gamma)^{-1} \right]_{ii}. \tag{E.37}$$

Since

$$\beta = \theta\Big(\frac{\phi_0}{\phi_1} + \frac{\alpha\kappa}{\sigma_\eta}\Big)\,\xi, \quad \frac{\partial\beta}{\partial\xi} > 0,$$

differentiating (E.37):

$$\frac{\partial V_{\delta_i,t}}{\partial\xi} = \zeta\,\Big[(\hat{\rho}I - \beta\Gamma)^{-1}\,\Gamma\,(\hat{\rho}I - \beta\Gamma)^{-1}\Big]_{ii}\,\frac{\partial\beta}{\partial\xi} > 0.$$

Thus, a decrease in ξ lowers $V_{\delta_i,t}$. Recall from (12) we have

$$-g_x\big(x_{i,t}, \lambda_{i,t}\big) = \theta\, V_{\delta_i,t}, \quad g_x < 0.$$

By the implicit-function theorem, $\partial x_{i,t}/\partial V_{\delta_i,t} > 0$. Therefore, a lower $V_{\delta_i,t}$ implies a lower $x_{i,t}$. From the dynamics of (E.36) we have

$$\frac{\partial\mathbb{E}[dx_{i,t}]}{\partial x_{j,t}} = \frac{\phi_0}{\phi_1}\,\gamma_{ij}\,\xi\,\theta, \quad \frac{\partial^2\mathbb{E}[dx_{i,t}]}{\partial x_{j,t}\,\partial\xi} = \frac{\phi_0}{\phi_1}\,\gamma_{ij}\,\theta > 0.$$