Third-Party Cookies, Data Sharing, and Return Comovement*

Si Cheng
sicheng@cuhk.edu.hk
CUHK Business School
Chinese University of Hong Kong

Yupeng Lin
bizliny@nus.edu.sg
NUS Business School
National University of Singapore

Ruichang Lu
ruichanglu@gsm.pku.edu.cn
Guanghua School of Management
Peking University

Xiaojun Zhang
zxj@gsm.pku.edu.cn
Guanghua School of Management
Peking University

December 6, 2021

Abstract

Third-party cookies connect different firms and facilitate data sharing. We find that common shocks to investor attention via cookie networks translate into economically significant comovement in terms of financial information acquisition, retail trading, and stock returns. An identification test based on the enactment of the California Consumer Privacy Act confirms this causal link. Furthermore, the return comovement among data-sharing firms is more pronounced in consumer-related industries, for more frequently installed cookies, and in the presence of joint human search on EDGAR and retail buying. Our findings document a beneficial effect whereby online data sharing alleviates the limited attention of investors and enhances information diffusion.

Keywords: Data Sharing, Return Comovement, Attention Spillover, Information Acquisition, Retail Trading

JEL code: D83, G11, G12, G14

*We thank Zhuo Chen, Xundi Diao, Pengfei Han, Shiyang Huang, Frank Weikai Li, Clark Liu, Kenny Phua, Baolian Wang, Junhong Yang, Bohui Zhang, Yingguang Zhang, Yu Zhang, and the conference and seminar participants at the 2021 China Advanced Research in Finance Conference, the 2021 China Finance Review International Conference, the 2021 China Fintech Research Conference, the 2021 Financial Management Association Annual Meeting, the 2021 Greater Bay Area Finance Conference, Chinese University of Hong Kong (Shenzhen), Korea University, Peking University, and Singapore Management University for their helpful comments.

Third-Party Cookies, Data Sharing, and Return Comovement

Abstract

Third-party cookies connect different firms and facilitate data sharing. We find that common shocks to investor attention via cookie networks translate into economically significant comovement in terms of financial information acquisition, retail trading, and stock returns. An identification test based on the enactment of the California Consumer Privacy Act confirms this causal link. Furthermore, the return comovement among data-sharing firms is more pronounced in consumer-related industries, for more frequently installed cookies, and in the presence of joint human search on EDGAR and retail buying. Our findings document a beneficial effect whereby online data sharing alleviates the limited attention of investors and enhances information diffusion.

Keywords: Data Sharing, Return Comovement, Attention Spillover, Information Acquisition, Retail Trading

JEL code: D83, G11, G12, G14

"The world's most valuable resource is no longer oil, but data"

Sam Jossen, The Economist, 2017 May 6th

1. Introduction

Data have become the most important asset in the digital era. Economic goods are usually rivals, but data are nonrivals. Data can be used simultaneously by any number of economic agents without being diminished. For example, multiple firms can use different algorithms to process the same set of data at the same time to guide various economic decisions without reducing the amount of data (Jones and Tonetti (2020)). The high economic value of data incentivizes many companies to utilize their digital platforms to collect and commercialize individual data, leading to unprecedented growth in the amount of data about individual preferences, social networks, and political views.[1] An increasing number of studies employ theoretical tools to examine the optimal allocation of data control rights, shedding light on the welfare implications of the undersharing or oversharing of data between individuals and platform companies (e.g., Bergemann and Bonatti (2015); Choi et al. (2019); Jones and Tonetti (2020); Acemoglu et al. (2021); Bergemann et al. (2021)).[2] Unlike extant studies, this paper seeks to provide first-hand evidence on the implications of data sharing for capital markets.

We focus on cookie networks and investigate the extent to which data sharing via cookies can affect investor attention and return comovement, shedding light on the efficiency of the capital market. Specifically, companies not only collect information on their customers to provide customized services or more targeted advertisements but also allow data brokers and other companies to install cookies on their websites so they can obtain supplemental information to facilitate retargeting advertisements and behavioral advertising (Bergemann and Bonatti (2019); Murgia and Harlow (2019)).[3] As a result, firms with common cookies could reach out to the same set of tracked users,

[1] Data vendor sales revenue is expected to reach $10.1 billion by 2022, more than triple the $3.1 billion observed in 2017 (Ram and Murgia (2019)).

[2] The nonrival nature of data and their important positive externalities can lead to underinvestment in data sharing (Jones and Tonetti (2020)). Another stream of studies incorporates the privacy cost of data sharing and speaks to concerns regarding excessive data usage and diffusion (e.g., Bergemann and Bonatti (2015); Acemoglu et al. (2021)). Such concerns have led to more stringent regulations, such as the General Data Protection Regulation (GDPR) in the European Union and the California Consumer Privacy Act (CCPA) in the state of California in the United States.

[3] Much of this information is collected via cookies, which are designed to follow users across the internet and record their browsing histories in real time to build a detailed, robust profile for each user.

who receive ads on various products and services they might be interested in according to their revealed profiles over time and recent search activities. We define firms that are connected through the same cookie network as data-sharing firms.

Although online data collection and sharing primarily target consumers, the same attention shocks can affect investors given their dual nature as consumers (e.g., Keloharju et al. (2012); Liaukonytė and Žaldokas (2020)). For instance, when investors visit the website of one firm in a specific cookie network, customized popup advertisements could direct them to other data-sharing firms, leading to an upward shift in attention to previously overlooked data-sharing firms.

Within the framework of Barber and Odean (2008), attention affects individual investors' buying more than their selling. The underlying reason for this phenomenon is that when buying a stock, investors need to choose from the entire universe of common stocks. Investors with bounded rationality are not able to process and rank all available stocks; instead, they choose to purchase stocks that have recently caught their attention (Odean (1999)). In contrast, investors' attention is not as constrained when selling because in the presence of short-sale constraints, most individual investors can sell only the stocks that they already own, and they typically hold only a few stocks. Therefore, it is likely that common attention shocks lead to comovement in net buying activities and subsequent stock returns.

In addition, different types of investor attention could have different asset pricing implications. On the one hand, since investors' cognitive capacity and attention are constrained, readily available information cannot be promptly incorporated into asset prices (e.g., Sims (2003); Peng and Xiong (2006); DellaVigna and Pollet (2009); Hirshleifer et al. (2009)). Data sharing via cookie networks could thus be meaningful in terms of attracting investor attention, encouraging information acquisition and subsequent trading, and improving pricing efficiency ("rational" attention). On the other hand, investors might overreact to stale public information and move prices away from fundamentals in the short run (e.g., Ho and Michaely (1988); Huberman and Regev (2001); Da et al. (2011); Tetlock (2011); Gilbert et al. (2012); Chawla et al. (2016)). This force could lead to negative capital market consequences of data sharing ("behavioral" attention). As a result, the effect of online data sharing on asset prices remains an empirical question.

To test the above hypotheses, we manually collect up-to-date cookie information from the websites of all listed U.S. companies. We first show that the daily stock returns of data-sharing

firms comove significantly with each other when common risk factors are controlled. Economically, a 1% increase in the data-sharing firms' return is associated with a daily abnormal return of 0.27% for a focal firm in the same cookie network. We further include lagged returns and do not detect any reversal effect, which suggests that the cookie network enhances the information diffusion between data-sharing firms rather than imposing a temporary price impact and shifts the equity demand curve (Merton (1971)). Our findings are also robust to controlling for a comprehensive set of pairwise firm characteristics that proxy for fundamental similarity.

We note that firms do not randomly join a data-sharing network. Unobservable firm characteristics may simultaneously affect the data-sharing decision and stock return comovement. To alleviate this concern, we rely on the enactment of the California Consumer Privacy Act of 2018 (CCPA) as an exogenous shock to the effectiveness and intensity of data sharing. The CCPA increases the hurdle for firms to collect and share individual information and therefore reduces data sharing. To the extent that California firms face greater litigation threats due to the enactment of the CCPA, we examine stocks with headquarters in California (i.e., the treatment group) and compare them to stocks with headquarters outside California (i.e., the control group). Using a standard difference-in-differences (DiD) setting based on the CCPA, we find a 41% reduction in return comovement between the focal firm and California data-sharing firms in the post-CCPA period. In contrast, the return comovement between the focal firm and non-California data-sharing firms remains unchanged. Our identification test supports a causal effect of data sharing on return comovement and documents that less data sharing significantly reduces return comovement.

To corroborate our argument, we conduct several cross-sectional tests. First, we utilize the heterogeneous effects of cookies on different industries. The targeted advertising function of cookies is more widely used in consumer-related industries. Therefore, we expect to find stronger return comovement among data-sharing firms in consumer-related industries than in other industries. Second, the nonrival nature of data leads to a strong economy of scope: the cookies used by more firms allow the platform company to provide more accurate profiling and better gauge investor attention. As such, data sharing via more frequently installed cookies could lead to more pronounced attention spillovers and hence stronger return comovement. The results of the cross-sectional tests support these predictions.

To provide more direct evidence on the underlying mechanisms driving this return comovement,

we utilize the log files from U.S. Securities and Exchange Commission (SEC)'s Electronic Data Gathering, Analysis, and Retrieval (EDGAR) system to identify investors' information acquisition activities (e.g., Drake et al. (2012, 2015); Lee et al. (2015); Drake et al. (2017); Ryans (2017)). These log files allow us to examine whether the search for a focal firm's financial information comoves with that of other data-sharing firms in the same cookie network. First, the focal firm's EDGAR search mostly comoves with that of data-sharing firms rather than the whole market. Second, human searches for the focal firm's information generally comove with human searches but not machine downloads of the data-sharing firms' information. These findings collectively provide direct support for our hypothesis that data sharing within the same cookie network leads to a common shock to investor attention and causes comovement in information acquisition. Such a rational response to attention shocks also echoes our previous findings showing no return reversal.

Second, we provide direct evidence on how data sharing affects retail trading activities, especially retail buying. We consider two sets of proxies for retail trading: one follows the algorithm proposed by Boehmer et al. (2021), identifying retail trades based on intraday trading data and computing retail order imbalance, and the other is based on changes in the number of Robinhood users. We find a uniformly significant comovement in retail trading among data-sharing firms. Importantly, this comovement nearly doubles when retail investors are net buyers of the data-sharing firms rather than net sellers.

To close the inferential loop, we further show stronger return comovement when firms display more correlated EDGAR search and correlated retail trading. Our findings are concentrated on EDGAR searches by human investors but not machine downloads and on retail investors who are net buyers but not net sellers. Taken together, these results suggest that online data sharing could lead to return comovement through the correlated financial information acquisition and subsequent buying of retail investors.

Finally, if data sharing causes return comovement through attention and information spillover, we can link data sharing to predictable variation in returns. Specifically, if a focal firm's own price declines while other data-sharing firms' price increases, we conjecture that the focal firm is relatively undervalued and that its price should increase once investors learn about the firm. The data-sharing portfolio return provides a reasonable benchmark for evaluating whether the focal firm is undervalued or overvalued. Relying on this implication, we develop a trading strategy to

exploit the lead-lag return predictability induced by information diffusion. We long the high-data-sharing-portfolio-return and low-own-return stocks and short the low-data-sharing-portfolio-return and high-own-return stocks. This long-short portfolio yields a daily return of 0.28% and a five factor-adjusted return of 0.27%. Consistent with the nature of instant attention shocks generated by the cookie network, the return predictability is short lived, as 88% (67%) of the 5-day (10-day) risk-adjusted return is concentrated on the first day. On the other hand, we do not find any subsequent reversal in stock returns, supporting the notion of information diffusion instead of a temporary price impact.

Overall, we show that online data sharing via cookies generates common attention shocks to data-sharing firms within the same network, resulting in a joint search for financial information and correlated trading activities. Enhanced investor attention and the subsequent trading accelerate the information diffusion process and lead to permanent price adjustments for firms in the same cookie network.

Our findings are related to several strands of the literature. We first enrich academic and policy discussions on the use of personal information. Existing studies focus on the market design and social externality of data sharing (e.g., Bergemann and Bonatti (2015, 2019); Acquisti et al. (2016); Easley et al. (2018); Choi et al. (2019); Jones and Tonetti (2020); Acemoglu et al. (2021); Bergemann et al. (2021); Cong et al. (2021); Liu et al. (2021)). The majority of extant studies focus on the excessive data sharing problem, providing interesting theoretical predictions but little empirical support. We instead focus on the nonrival nature of data and the resultant beneficial effect of data sharing. Our analyses provide first-hand empirical evidence on how data sharing can mitigate the limited attention problem in the capital market. The explosion of digital footprints and technological innovations create granular segments of consumer characteristics; hence, firms can easily target consumers with similar attributes and preferences. Our findings imply that personalized ads also cater to the interests of investors, leading to more financial information search and stock return comovement.[4]

Second, our paper contributes to the literature on investor attention and individual trading behaviors (e.g., Odean (1998); Peng and Xiong (2006); Barber and Odean (2008); Da et al. (2011);

[4]Note that we do not aim to assess the overall welfare implications of data sharing; instead, we focus on the capital market consequences of the cookie network and document a potential beneficial effect.

Tetlock (2011); Abel et al. (2013); Chen et al. (2021)) and information acquisition (e.g., Drake et al. (2012, 2015); Lee et al. (2015); Drake et al. (2017); Ryans (2017); Blankespoor et al. (2019, 2020)).[5] Unlike traditional information distribution methods, online data sharing allows for more dynamic, interactive, and personalized information dispersal and therefore is more likely to capture investors' attention. More important, shocks to investor attention are instantly generated in a highly integrated information market, providing a unique testing ground to analyze the spillover effect of investor attention induced by data sharing. We follow the growing body of literature that measures information acquisition using EDGAR, providing direct evidence to support information acquisition comovement due to common shocks to investor attention via cookie networks.[6] We further utilize a newly developed algorithm, namely, that of Boehmer et al. (2021), to identify retail trades and a novel dataset on Robinhood users to provide direct evidence on attention-driven comovment in the context of retail trading. Our paper shows that online data sharing alleviates the limited attention of investors and their underreaction to news, highlighting that targeted attention shocks enhance information acquisition and help incorporate new information into stock prices.

We also contribute to the literature on stock return comovement and particularly to studies exploring the role of information diffusion. Past work links comovement to analyst coverage (Muslu et al. (2014); Hameed et al. (2015)) and common underwriters (Grullon et al. (2014)). We instead emphasize the role of online data sharing, which relies on mechanisms other than traditional information intermediaries to facilitate information spillover. Our findings also relate to recent work on investor attention and comovement (e.g., Drake et al. (2017); Huang et al. (2019); Jiang et al. (2019); Chen et al. (2021)). Our novelty is to illustrate the spillover effect on a specific set of data-sharing stocks (i.e., through common cookies) beyond comovement with industry and market peers and provide direct evidence on the economic mechanism, i.e., investor attention is translated into more information acquisition and subsequent return comovement.

The remainder of this paper is organized as follows. Section 2 describes the data and the main variables used and presents some stylized characteristics associated with data sharing. Section

[5]Our paper is also related to studies exploring how product advertising affects investor attention and information acquisition (e.g., Keloharju et al. (2012); Lou (2014); Madsen and Niessner (2019); Focke et al. (2020); Liaukonytė and Žaldokas (2020); Mayer (2021)).

[6]See, e.g., Drake et al. (2012), Drake et al. (2015), Lee et al. (2015), Dechow et al. (2016), Drake et al. (2016), Bozanic et al. (2017), Drake et al. (2017), Ryans (2017), Li and Sun (2019), Bernard et al. (2020), and Liaukonytė and Žaldokas (2020).

3 relates return comovement to data sharing. Section 4 analyzes comovements in information acquisition and retail trading due to data sharing. Section 5 develops a trading strategy that exploits information diffusion in cookie networks and presents additional analyses. A brief conclusion follows in Section 6.

2. Data and Main Variables

Our sample includes all common stocks trading on NYSE/AMEX/Nasdaq from 2015 to 2019 due to the availability of data sharing measures.[7] We obtain daily and monthly stock data from the Center for Research in Security Prices (CRSP). Quarterly and annual financial statement data come from the COMPUSTAT database. Analyst forecast data come from the Institutional Brokers' Estimate System (I/B/E/S). We acquire quarterly institutional equity holdings from the Thomson-Reuters Institutional Holdings (13F) database.[8] We also obtain the server request records from the SEC EDGAR Log File dataset, which contains all internet search traffic for SEC filings.[9]

2.1. Measuring Data Sharing

2.1.1. What is a Cookie?

A cookie (also called a web cookie, Internet cookie, browser cookie) is a small piece of data (i.e., tracking code) stored on the user's computer when browsing a website. Cookies are designed to be a reliable mechanism for websites to remember information (such as items added to the shopping cart in an online store and language preference) or to record the user's browsing activity (such as clicking particular buttons, logging in, and past website visits). They can also be used to remember pieces of information that the user previously entered into form fields, such as names, addresses, passwords, and payment details. Cookies are useful—they allow modern websites to work the way people have come to expect—with an increasing level of personalization and rich interactive

[7]Since we only have a snapshot of the usage of cookies in April 2020, we use the latest five years as our testing sample. Later, we perform a robustness test using only data from 2019 and find similar results.

[8]The institutional ownership data come from money managers' quarterly 13F filings with the SEC. The database contains the positions of all institutional investment managers with more than $100 million U.S. dollars under discretionary management. All holdings worth more than $200,000 U.S. dollars or 10,000 shares are reported in the database.

[9]The data are available at `https://www.sec.gov/dera/data/edgar-log-file-data-set.html`. All tests using the EDGAR data end in June 2017 due to data availability.

functionality.[10]

If the host domain for a cookie is different from the company's domain, it is a third-party cookie. They are usually placed on a website via scripts or tags added to the webpage and placed by cookie platforms that work with many companies and play an intermediary role as data brokers. As shown in Figure 1, 10 cookies are placed on Verizon's website (highlighted in the bottom-left corner), among which 8 are third-party cookies and belong to 4 different platforms (Doubleclick/Google, Demdex/Adobe, Contentsqure, and lpsnmedia/LivePerson).

Third-party cookies could also bring additional functionality to the site, such as enabling content to be shared via social networks.[11] More important, third-party cookies are widely used for retargeting advertisements and behavioral advertising. By adding tags to a page, advertisers can track users or their devices across different websites. When users visit another site with the same tag, it reports to the advertiser the site they were last on when the cookie was set. By aggregating the information across millions of visits on different sites, it enables the advertiser to develop a detailed, robust profile for each user through her browsing history.[12] The advertiser then uses this information to display more personalized and targeted advertisements based on the user's perceived habits and interests as well as recent search activities.

2.1.2. Data Sharing among Firms

We use the common third-party cookies shared between firms to measure firms' data sharing activities. Third-party cookies are placed by cookie platforms that work with many companies, which aim to create customized advertisements based on an individual's personal preferences and promote various products and services. Customized popup advertisements could generate attention shocks for internet users, and firms that share the same third-party cookies are likely to be jointly recommended. As a result, the user might be directed to other data-sharing firms due to common third-party cookies. Therefore, third-party cookies foster a data-sharing network within which the data collected from one company could facilitate the marketing activities of another company.

[10] For perspective, Daily Mail and The Telegraph have 19,136 and 14,025 third-party cookies on their sites, respectively (Davies (2017)).

[11] For instance, a website owner could use a piece of code provided by YouTube and include a YouTube video on its webpage. YouTube will then be able to set cookies through this code and collect information on users' browsing activity.

[12] Although third-party cookies mostly target the browser rather than the user, as most people log in and use the same browser regularly, the digital footprints collected over time can be highly personalized.

We obtain up-to-date cookie information from the website of the company in April 2020. First, we obtain the set of listed firms with available accounting information and company website addresses in fiscal year 2019 from COMPUSTAT. Second, we use OpenWPM to crawl all the cookies from each company's website. OpenWPM is a web privacy measurement framework that allows researchers to collect data on a scale of thousands to millions of websites (Englehardt and Narayanan (2016); Ramadorai et al. (2019)). Third, we trace back the ultimate owner (i.e., the platform such as Google Ads and Adobe Audience Manager) of the third-party cookies using Cookiepedia. Cookiepedia is the largest database of pre-category cookies and online tracking technologies.[13] Firms that use third-party cookies from the same platform, e.g., Google Ads, share the internet traffic data collected from their websites with each other. In this way, for each focal firm, we identify the group of data-sharing firms that adopt the same third-party cookies. Figure 2 illustrates the conceptual framework of data sharing through cookies and the definitions of data-sharing firms and non-data-sharing firms.

One limitation is that we do not have full time series for the cookies. However, we conduct a validation test by crawling the cookies again in October 2020 and find that the total number of cookies (including first-party and third-party cookies) increases by 0.86%. Thus, the cookie network is likely to be stable over time. In addition, firms are unlikely to opt into a cookie network due to expected return comovement or attention comovement in the future; therefore, we are less concerned about this data limitation.

2.2. Measuring Excess Return Comovement

For each focal firm i that has at least one third-party cookie, we construct a *data-sharing portfolio* that consists of firms with common third-party cookies (i.e., firms with third-party cookies from the same platform). The firms that are classified into the data-sharing portfolio are held constant in our analyses.

To measure the excess return comovement with a data-sharing portfolio, we regress stock returns on the data-sharing portfolio returns, controlling for the effects of common risk factors. Specifically,

[13]The Cookiepedia site was established to fill a gap in information about what cookies do, who is using them for what purposes, and how to manage cookies and is maintained by OneTrust, a privacy management software company.

we estimate the following daily regression model:

$$R_{i,d} = \alpha_0 + \beta_1 DSRET_{i,d} + \gamma_1 MKT_d + \gamma_2 SMB_d + \gamma_3 HML_d + \gamma_4 MOM_d + \epsilon_{i,d}, \quad (1)$$

where $R_{i,d}$ is the excess return of stock i on day d and $DSRET_{i,d}$ is the (equal-weighted) excess return of stock i's data-sharing portfolio. We further adjust for the common risk factors based on the Fama-French-Carhart (FFC) four-factor model consisting of the market factor (MKT, defined as the excess return on the value-weighted CRSP market index over the one-month Treasury bill rate), the size factor (SMB, defined as small minus big firm return premium), and the book-to-market factor (HML, defined as the high book-to-market minus the low book-to-market return premium) (Fama and French (1993)), and the Carhart (1997) momentum factor (MOM, defined as the winner minus loser return premium).[14]

We also consider several alternative model specifications. First, we replace $DSRET_{i,d}$ in Equation (1) with its residual based on the FFC four-factor model, following Hameed and Xie (2019). Specifically, we regress $DSRET_{i,d}$ on the FFC four factors over a one-year estimation period to obtain betas of a stock. The residual (denoted as $ResidualDSRET_{i,d}$) is computed as the realized stock return minus the product of the stock's lagged four-factor betas and the realized four-factor returns of a given day. Second, we control for the lagged returns of stock i's data-sharing portfolio. Finally, we extend to pairwise regressions and further control for a host of pair characteristics, following Antón and Polk (2014).

2.3. Descriptive Statistics

We start with 4,732 unique firms with available cookie information. We then merge these data with platform information from Cookiepedia, and identify 2,274 unique firms (9,142 unique firm-platform pairs) that have at least one third-party cookie. Since data sharing relies on common third-party cookies, our final sample includes 1,558 unique data-sharing firms (7,083 unique firm-platform pairs).

Table 1 presents a list of top platforms that own third-party cookies. For each platform, we report the number of unique industries and the number of unique firms that adopt the platform's

[14]We thank Kenneth French for making the common factor returns available via his website: `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`.

third-party cookie. Industry is defined by 2-digit SIC code. We find that different industries use the same set of platforms. For example, Google's cookie is used by 62 different industries, suggesting that cookie usage is not industry specific. We further illustrate the data structure by providing firm-level statistics. We find that the cookie market is dominated by several large platforms. For instance, cookies from Google and Facebook are installed by 959 and 741 firms, respectively, while cookies from Yahoo (ranked 10^{th}) are installed by only 141 firms in our sample.

To further investigate whether some platforms specialize in a few industries, we calculate the industry concentration for each platform in the last column. We find that the industry concentration—defined as the Herfindahl-Hirschman Index (HHI) based on the number of adopters in each industry—is low, and only one platform (i.e., Demandbase) has an HHI larger than 0.15. Overall, we find that third-party cookies are adopted by a wide range of industries and firms, consistent with the tremendous size of the online data market.

Panel A of Table 2 reports the summary statistics for a list of firm characteristics in 2019. There are 1,348 unique firms with at least one third-party cookie and available accounting information. The average firm adopts 4.38 third-party cookies. The cross-sectional variation is substantial, as the standard deviation is 5.28. In addition, the overall characteristics of the firms are comparable to those of the average publicly listed firms in the universe.

In Panel B of Table 2, we independently sort stocks according to the number of third-party cookies adopted by the firm and quintile of firm size in 2019. We find that 52.2% of the firms (1,471 out of 2,819) do not place any third-party cookies on their websites, while the remaining 47.8% of the firms (1,348 out of 2,819) allow for third-party cookies. Among the firms with third-party cookies, 10.3% of the firms (139 out of 1,348) place 10 or more third-party cookies on their websites. The number of firms with a given number of third-party cookies is evenly distributed across each size quintile in most cases. Thus, we conclude that there is no correlation between firm size and firms' decision to place third-party cookies on their websites. Panel C reports similar statistics when we replace firm size with the market-to-book ratio. We find that cookie adoption is also unlikely to be correlated with the market-to-book ratio.

3. Data Sharing and Return Comovement

3.1. The Baseline Analysis of Return Comovement

We start with the baseline analysis described in Equation (1) to examine the return comovement within the full sample. The standard errors are clustered by calendar day to account for cross-correlation in stock returns. Panel A of Table 3 reports the results. We find that *DSRET* is highly significant with a coefficient of 0.266 (t-statistic = 15.91), as shown in Model 1. The effect is also economically large, i.e., a 1% increase in the data-sharing firms' return is associated with an abnormal return of 0.27% per day for the focal firm. This suggests that data sharing generates excessive comovement among firms after controlling for the common risk factors.

In Model 2, we replace $DSRET_{i,d}$ in Equation (1) with its residual based on the FFC four-factor model as described above, i.e., $ResidualDSRET_{i,d}$. The results are qualitatively and quantitatively the same, and the focal firm tends to comove with other data-sharing firms.

We further include lagged returns of the data-sharing portfolio in Model 3. Specifically, we include lagged 1-day, 2-day, 3-day, 4-day, and 1-week (day $d-9$ to $d-5$) returns of the data-sharing portfolio. We find that the coefficients of all lagged terms are much smaller than the contemporaneous term, i.e., 0.028 for the lagged 1-day return vs. 0.255 for the contemporaneous return. The magnitudes of the coefficients on lagged returns indicate that the lagged effect is not economically meaningful. More important, we do not find any reversal in past performance, implying that the price adjustment associated with data sharing is permanent. This is consistent with the notion that the cookie network enhances the information diffusion between data-sharing firms rather than imposing a temporary price impact.

Model 4 further includes industry fixed effects to control for time-invariant industry characteristics, and our findings remain unchanged.[15] For instance, a 1% increase in the data-sharing firms' return is associated with an abnormal return of 0.27% per day for the focal firm.

One caveat regarding our analysis is that cookie adoption is measured by a snapshot in April 2020. The statistical argument we make here is that any noise in the classification of data-sharing firms could lead to an underestimation of the comovement effect. As a robustness check, we repeat

[15]The industry fixed effects are defined based on 4-digit SIC codes. Unreported results are also robust to controlling for both industry and calendar day fixed effects.

the analysis in Panel A based on the most recent year—2019. The cookie network we capture should be very close to the real cookie network in 2019. As shown in Panel B of Table 3, we find consistent evidence of return comovement among data-sharing firms, i.e., a 1% increase in the data-sharing firms' return is associated with an abnormal return of 0.29% per day for the focal firm (Model 1). Our findings are also robust to all regression specifications. Overall, our baseline results show that online data sharing plays an important role in explaining stock return comovement. These preliminary findings support the information diffusion hypothesis, and we provide additional evidence to analyze the economic mechanism in later sections.

We note that one non-mutually exclusive explanation is that correlated cash flows among data-sharing firms could drive stock return comovement. To minimize the cash flow effect, we explicitly control for the common risk factors and focus on the daily stock returns, as they are less likely affected by the underlying cash flow comovement. In later analyses, we explore an identification test that plausibly exogenously reduces data sharing to establish a causal relationship. We also investigate the cross-sectional variation in firm characteristics and provide additional evidence when controlling for similarity in a comprehensive set of firm and industry characteristics. To further support daily return comovement, we examine the economic drivers related to instant attention shocks and daily information acquisition and trading activities.

3.2. Identification Test

One concern about our main specification is that firms may not randomly join a data-sharing network. Unobservable firm characteristics may simultaneously affect the data-sharing decision and stock return comovement. In this subsection, we utilize the enactment of the CCPA as a laboratory test that plausibly exogenously reduces data sharing to establish a causal relationship. The CCPA was introduced in 2018 and gives consumers more control over the personal information that businesses collect about them.[16] This law enhances privacy rights and consumer protection for residents of California, including (1) the right to know about the personal information a business collects about them and how it is used and shared; (2) the right to delete personal information collected from them; (3) the right to opt out of the sale of their personal information; and (4) the right to nondiscrimination for exercising their CCPA rights. The enactment of CCPA increases

[16]See the State of California Department of Justice website for details: `https://oag.ca.gov/privacy/ccpa`.

the hurdle of data collection and data sharing via cookies and therefore should reduce attention comovement and return comovement. It is also reasonable to believe that the adoption of CCPA is not driven by individual firm characteristics, satisfying the exclusion condition.

We note that almost all large firms doing business in California are affected by the CCPA.[17] To the extent that the enforcement of the CCPA is constrained by resources of the California Attorney General, we assume that the litigation risk is higher for firms with headquarters in California than for other firms. The underlying rationale is that public enforcement strength is negatively correlated with geographic distance (Kedia and Rajgopal (2011)). In addition, any noise in the data (e.g., non-California firms are similarly affected by the CCPA) would reduce the difference between California and non-California firms and bias against finding significant results.

To proceed, our identification strategy involves examining stocks with headquarters in California (i.e., the treatment group) and comparing them to stocks with headquarters outside California (i.e., the control group). We then separate the data-sharing portfolio into two portfolios based on headquarters location and conduct a standard DiD estimation via daily regression:

$$\begin{aligned} R_{i,d} = \alpha_0 + \beta_1 DSRET\_CA_{i,d} + \beta_2 DSRET\_non\text{-}CA_{i,d} + \beta_3 DSRET\_CA_{i,d} \times Post_d \\ + \beta_4 DSRET\_non\text{-}CA_{i,d} \times Post_d + \beta_5 Post_d + \boldsymbol{\gamma}' \boldsymbol{F}_d + \epsilon_{i,d}, \end{aligned} \tag{2}$$

where $DSRET\_CA_{i,d}$ and $DSRET\_non\text{-}CA_{i,d}$ are the excess return of stock i's data-sharing portfolio with headquarters inside and outside California on day d, respectively.[18] $Post_d$ represents several dummy variables: $Post\ 2Y$ equals 1 for two years after the introduction of the CCPA (i.e., 2018–2019) and 0 otherwise (i.e., 2015–2017); $Post^{+1}$ equals 1 for one year after the CCPA (i.e., 2018) and 0 otherwise; and $Post^{+2}$ equals 1 for the second year after the CCPA (i.e., 2019) and 0 otherwise. Vector $\boldsymbol{F}$ stacks the FFC four factors. The standard errors are clustered by calendar day.

[17] The CCPA applies to for-profit businesses that do business in California and meet any of the following criteria: (1) have a gross annual revenue of over $25 million; (2) buy, receive, or sell the personal information of 50,000 or more California residents, households, or devices; or (3) derive 50% or more of their annual revenue from selling California residents' personal information.

[18] To ensure that our estimates are not driven by the different numbers of firms with headquarters inside and outside California, we scale both portfolios by the total number of firms. Specifically, $DSRET\_CA_{i,d} = \frac{\sum_{i \in CA} R_{i,d}}{N_d}$, $DSRET\_non\text{-}CA_{i,d} = \frac{\sum_{i \notin CA} R_{i,d}}{N_d}$, where $R_{i,d}$ is the excess return of stock i on day d, and N_d is the total number of firms. $i \in CA$ and $i \notin CA$ indicate that stock i has headquarters inside and outside California, respectively. By construction, $DSRET\_CA_{i,d} + DSRET\_non\text{-}CA_{i,d} = DSRET_{i,d}$ in Equation (1).

We focus on $\beta_3 - \beta_4$ in Equation (2), as it captures the change in return comovement for the treatment group relative to the control group for the post-CCPA period (compared to the pre-CCPA period). We expect that the comovement between the focal firm and the California-data-sharing portfolio is weaker than that between the focal firm and the non-California-data-sharing portfolio during the post-CCPA period, i.e., $\beta_3 - \beta_4 < 0$.

The results are reported in Table 4. Model 1 estimates a simplified version of Equation (2) to demonstrate a general relationship, i.e., the focal firm tends to comove slightly less with its California-data-sharing portfolio, possibly due to more stringent privacy regulation. More important, using the introduction of the CCPA as an exogenous shock to the intensity of data sharing, we find that the return comovement between the focal firm and California-data-sharing firms significantly declines by 0.165 in the post-CCPA period, accounting for 41% ($-0.165/0.403$) of the return comovement in the pre-CCPA period (Model 2). In contrast, the return comovement between the focal firm and non-California-data-sharing firms does not change in the post-CCPA period. This finding suggests that the enactment of the CCPA only affected California-data-sharing firms, further justifying the validity of the experiment. Importantly, the DiD estimate (i.e., $\beta_3 - \beta_4$) amounts to -0.175 and is statistically significant at the 5% level (F-statistic = 5.35).

Model 3 further investigates how the negative effect on return comovement evolves over time after the CCPA. The CCPA was signed into law on June 28, 2018, and became effective on January 1, 2020. Since it takes time for firms to adapt to new regulation, we expect the reduction in return comovement to be stronger in 2019 than in 2018. Consistent with our conjecture, we find that the return comovement between the focal firm and the California-data-sharing firms does not decline significantly in 2018 but weakens considerably in 2019, accounting for 58% of the return comovement in the pre-CCPA period ($-0.234/0.406$). In addition, we do not find a significant change in the return comovement between the focal firm and the non-California-data-sharing firms in either year. The DiD estimate is -0.075 and statistically insignificant (F-statistic = 0.6) for 2018 and -0.255 and statistically significant at the 1% level (F-statistic = 7.14) for 2019. The dynamic pattern provides additional supportive evidence for a causal relationship between data sharing and return comovement. Our findings are also robust to including industry fixed effects as shown in Models 4-6.

Overall, our identification test explores plausible exogenous variations in the effectiveness and

intensity of data sharing and provides evidence to support the causal effect of data sharing on return comovement. When data sharing is reduced due to the enhanced privacy rights following the enactment of the CCPA, return comovement declines significantly.

3.3. Heterogeneity in Data Sharing

Next, we investigate the cross-sectional variation in data sharing. Since the primary purpose of third-party cookies is to facilitate retargeting advertisements and behavioral advertising to directly generate sales, we expect firms in consumer-related industries to comove more with the data-sharing portfolio than firms in other industries.[19]

The results are tabulated in Panel A of Table 5. We report the estimates of Equation (1) based on the subsamples, with Models 1-3 for firms in consumer-related industries and Models 4-6 for other industries. Consistent with our conjecture, a focal firm in a consumer-related industry displays higher comovement with other data-sharing firms than with a focal firm in another industry. For instance, a 1% increase in the data-sharing firms' return is associated with an abnormal return of 0.43% (0.18%) per day for the focal firm in a consumer-related industry (other industry) in Model 1 (Model 4).[20] Our findings remain qualitatively and quantitatively similar under alternative regression specifications.

Due to the economy of scope and the network effect, the cookies used by more firms are more central in the network and more likely to collect information on an individual user; thus, they build a granular profile for each user and facilitate more efficient data sharing.[21] Therefore, we conjecture that the adoption of high-frequency cookies is associated with more return comovement. Specifically, we separate the third-party cookies into two groups based on their usage frequency—cookies owned by the top 10 platforms in Table 1 are classified as high-frequency cookies, and the remaining cookies are classified as low-frequency cookies.

We repeat the analysis of Equation (1) and report the results in Panel B of Table 5, with

[19]Specifically, we define consumer-related industries using the two-digit SIC code. Firms with two-digit SIC codes of 01-09 (Agriculture, Forestry, and Fishing), 40-49 (Transportation, Communications, Electric, Gas, and Sanitary Services), 52-59 (Retail Trade), 60-67 (Finance, Insurance, and Real Estate), and 70-89 (Services) are defined as consumer related.

[20]Unreported results confirm that our findings are robust to alternative industry definitions, i.e., inclusion or exclusion of industries other than retail trade.

[21]A cookie platform is "central" in the cookie network if it is adopted by many firms, or in other words, has many direct connections with other firms in the network as measured by its degree of centrality (e.g., Freeman (1977); El-Khatib et al. (2015)).

Models 1-3 for high-frequency cookies and Models 4-6 for low-frequency cookies. We find that the focal firms adopting high-frequency cookies display considerably more comovement with data-sharing portfolios than those adopting low-frequency cookies. For instance, a 1% increase in the data-sharing firms' return is associated with an abnormal return of 0.55% (0.18%) per day for a focal firm with high-frequency (low-frequency) cookies in Model 1 (Model 4). Our findings remain unchanged in alternative regression specifications.

An untabulated test further divides the sample into two groups based on total institutional ownership. We find that the return comovement between the focal firms and data sharing firms is weaker among the firms with high institutional ownership; this is consistent with the notion that third-party cookies mostly attract the attention of retail investors. Later, we will provide more direct evidence on retail investors' financial information acquisition and trading activities.

Collectively, we find that return comovement is one to two times higher when data sharing is more important (e.g., consumer-related industries) and more effective (e.g., high-frequency cookies), reinforcing the link between data sharing and return comovement.

3.4. Pairwise Analysis

To further alleviate the concern that the data sharing decision may coincide with other firm fundamentals, we perform regression analysis to control for a comprehensive set of firm characteristics, following Antón and Polk (2014). Specifically, we estimate the following monthly Fama and MacBeth (1973) regression:

$$ARCORR_{ij,t} = \alpha_0 + \beta_1 DS_{ij,t-1} + \boldsymbol{\gamma}' \boldsymbol{N}_{ij,t-1} + \epsilon_{ij,t}, \tag{3}$$

where $ARCORR_{ij,t}$ is the correlation of daily FFC four-factor abnormal returns between stocks i and j in month t. $DS_{ij,t-1}$ refers to a list of data sharing variables for each stock pair: $NUMTP$ is the number of common third-party cookies; $LOGNUMTP$ is the logarithm of one plus the number of common third-party cookies; $\%NUMTP$ is the percentage of common third-party cookies; and $DNUMTP$ is a dummy variable that equals 1 if $NUMTP > 0$ and 0 otherwise. Vector $\boldsymbol{N}$ stacks all other control variables for each stock pair, including $FCAP$, defined as the total value of stock held by common funds investing in both stocks, scaled by the total market capitalization of the

two stocks; *NUMANA*, defined as the number of analysts that issued at least one annual earnings forecast for both stocks during the previous 12 months; *SAMESIZE*, *SAMEBM*, and *SAMEMOM*, defined as the negative of the absolute difference in percentile ranking for market capitalization, book-to-market ratio, and past 12-month return across a pair, respectively; *NUMSIC*, defined as the number of consecutive SIC digits, beginning with the first digit, that are equal for a pair; *SIZE1* and *SIZE2*, defined as the normalized rank-transform of the percentile market capitalization of the two stocks, as well as the interaction between them *SIZE1* × *SIZE2*; *RETCORR*, *ROECORR*, and *VOLCORR*, defined as the correlation in the two stocks' past return, return on equity, and abnormal trading volume, respectively; *DIFFGROWTH*, *DIFFLEV*, and *DIFFPRICE*, defined as the absolute difference in the two stocks' log sales growth rate, financial leverage ratio, and log share price, respectively; and *DSTATE*, *DINDEX*, and *DLISTING*, defined as a dummy variable that equals 1 if the two firms are located in the same state, belong to the S&P 500 index, and are listed on the same stock exchange, and 0 otherwise, respectively. We also create additional controls based on the normalized rank transform of the percentile book-to-market ratio (past 12-month return) of the two stocks and the interaction between them, i.e., *BM1*, *BM2*, and *BM1* × *BM2* (*MOM1*, *MOM2*, and *MOM1* × *MOM2*), and further control for the nonlinear relationship captured by $SAMESIZE^2$, $SAMESIZE^3$, $SAMEBM^2$, $SAMEBM^3$, $SAMEMOM^2$, and $SAMEMOM^3$. Appendix A provides the detailed definition of each variable. We also report Newey and West (1987) adjusted t-statistics.

We tabulate the results in Table 6. We find that data sharing is positively associated with the four-factor residual correlation in stock returns after controlling for a comprehensive set of pairwise firm characteristics that proxy for fundamental similarity. This result holds across different specifications both statistically and economically. In particular, a one-standard-deviation increase in *NUMTP* is associated with a 0.19% higher residual stock return comovement in Model 4, which translates into a 52% increase relative to the average comovement of 0.365% for the sample.[22] Our findings are robust to alternative proxies for data sharing that address the potential skewness in the number of common third-party cookies. For instance, a one-standard-deviation increase in *LOGNUMTP* (*%NUMTP*) is associated with a 0.20% (0.12%) higher residual stock return comove-

[22]The impact of *NUMTP* is 0.19%, computed as 0.125% × 1.531, where 0.125 is the regression coefficient in Model 4, and 1.531 is the standard deviation of *NUMTP*.

ment in Model 6 (Model 8).[23] In addition, as shown in Model 10, stock pairs with common cookies display 0.28% higher residual return comovement than those without common cookies, accounting for a 76% increase relative to the sample average.

Since we control for common coverage by institutional investors and sell-side analysts, as well as the similarity in firm characteristics, our findings support the notion that online data sharing plays an incremental role in generating common attention shocks among retail investors and subsequent return comovement among firms in the same cookie network. Our results also confirm the dual nature of investors as consumers (e.g., Keloharju et al. (2012); Liaukonytė and Žaldokas (2020)), highlighting the spillover effect of consumer-oriented data collection and sharing on the financial market.

4. How Does Data Sharing Affect Return Comovement?

In the previous section, we show that stock returns of data-sharing firms within a cookie network exhibit strong comovement. In the following, we investigate the underlying mechanisms through which data sharing affects return comovement. We hypothesize that online data sharing induces correlated investor attention, resulting in common investment behavior. When investors search for a firm on the internet, other data-sharing firms in its cookie network are more likely to be seen due to retargeting advertisements and behavioral advertising. Since investor attention is a scarce resource especially in the context of making buying decisions, individual investors are net buyers of attention-grabbing stocks (Barber and Odean (2008)).[24] In this way, common third-party cookies could lead to correlated attention shocks and correlated net buying among data-sharing firms and subsequent return comovement. In this section, we examine whether data sharing leads to (1) correlated financial information acquisition and (2) correlated retail trading. We then investigate how data sharing, correlated information acquisition, and correlated retail trading jointly affect return comovement.

[23] The standard deviation of *LOGNUMTP* and *%NUMTP* in the full sample is 0.549 and 0.083, respectively.

[24] Investors' attention is not as constrained when they are deciding which stocks to sell, because individual investors typically hold only a few stocks and cannot easily short sell.

4.1. Comovement in Information Acquisition

We explicitly measure investors' attention based on their information acquisition activities via the EDGAR system. The EDGAR log files record the network (IP) address of each user that downloads a document. A large accounting and finance literature uses EDGAR logs to study the demand for financial information (e.g., Drake et al. (2012, 2015); Lee et al. (2015); Dechow et al. (2016); Drake et al. (2016); Bozanic et al. (2017); Drake et al. (2017); Ryans (2017); Li and Sun (2019); Bernard et al. (2020); Liaukonytė and Žaldokas (2020)). While investors can access SEC filings from other sources, such as a firm's investor relations website, Yahoo! Finance, and Bloomberg, it appears that EDGAR captures a significant fraction of financial disclosure demand.[25] More important, the majority of EDGAR users are likely to be individual investors (Li and Sun (2019)), and the acquisition of financial information is mostly driven by investment needs rather than consumption needs; thus, EDGAR search provides an ideal setting for us to test the effect of data sharing on retail attention and subsequent return comovement.

Following Ryans (2017), we first identify programmatic downloads (labeled robots) and consider the remaining page views as human search. The general robot-screening procedure calculates statistics about a user's download patterns over a day and then applies one or more tests to classify the user as a robot or a human. Specifically, (1) humans do not download more than 25 items in a single minute; (2) humans do not download more than 3 different companies' items in a single minute; and (3) humans do not download more than 500 items in a single day.

To measure the comovement in EDGAR search, we estimate the following daily regressions:

$$HUM_{i,d} = \alpha_0 + \beta_1 DSHUM_{i,d} + \beta_2 DSROB_{i,d} + \beta_3 MKTHUM_d + \beta_4 MKTROB_d + \epsilon_{i,d}, \quad (4)$$

where $HUM_{i,d}$ is the number of page views by human readers (i.e., human search) of stock i on day d; $DSHUM_{i,d}$ and $DSROB_{i,d}$ are the (equal-weighted) number of human searches and robot searches of stock i's data-sharing portfolio, respectively; and $MKTHUM_d$ and $MKTROB_d$ are the (equal-weighted) number of human searches and robot searches of the market portfolio, respectively. The standard errors are clustered by calendar day.

The results are reported in Table 7. First, as shown in Model 1, human search for focal firms is

[25] See Li and Sun (2019) for a discussion on the advantages of SEC EDGAR over other information sources.

highly correlated with human search for other data-sharing firms compared with the market average, i.e., 0.801 vs. 0.209. This result supports the argument that data sharing leads to attention comovment and joint search for financial information. The correlated human search among data-sharing firms is distinct from the market-wide search activities that could be driven by macroeconomic information. Second, Model 2 further controls for robot search measures. Our conjecture is that data sharing leads to comovement in investors' attention and should not affect robot search, which relies on computer programs to download large volumes of data instead of selectively searching for a few firms. We find consistent results that human search for the focal firms is highly correlated with human search but not robot search for the data-sharing firms, i.e., 0.717 vs. 0.021. Finally, our findings remain intact after controlling for industry fixed effects (Models 3-4).

As a robustness check (untabulated), we employ winsorized versions of the search measures, and the results largely remain unchanged.[26] Overall, we identify an important economic mechanism that drives return comovement, i.e., online data sharing generates correlated attention shocks and facilitates comovement in information acquisition.

Note that a joint EDGAR search per se does not imply that investors uncover value-relevant new information about the firms, as investors could still irrationally react to stale public information and move the price away from fundamentals (e.g., Ho and Michaely (1988); Barber and Loeffler (1993); Liang (1999); Huberman and Regev (2001); Da et al. (2011); Tetlock (2011); Engelberg et al. (2012); Gilbert et al. (2012); Chawla et al. (2016); Da et al. (2020); Barber et al. (2021); Chen et al. (2021)). Combining these results with our early findings showing no return reversal, we conclude that retail investors searching through the EDGAR system appear to be relatively sophisticated, and their information acquisition activities reveal new information. Our results also suggest that online data sharing within common third-party cookies differs from the role of other online platforms in gauging investor attention and information dissemination. For instance, the retail attention captured by Google search volumes (Da et al. (2011, 2020)) and news diffusion through Twitter (Chawla et al. (2016)) leads to transitory price pressure and then reversal, suggesting that investors overreact to stale news rather than incorporating new information into stock prices. This difference could be attributed to the more dynamic, interactive, and personalized nature of targeted ads, as they are

[26] We winsorize the top and bottom 1% of the data. Since the EDGAR page views are highly skewed, these alternative measures alleviate the concern that the results may be driven by certain outlier stocks.

more likely to cater to consumers and investors who are interested in the corresponding firms. As a result, investors respond more rationally to targeted attention shocks.

4.2. Comovement in Retail Trading

While financial information acquisition through EDGAR searches is likely to facilitate subsequent trading and price adjustments, in this section, we provide direct evidence on how data sharing affects retail trading activities. We consider two sets of proxies for retail trading.

First, we follow the algorithm proposed by Boehmer et al. (2021) to identify retail trades from TAQ data. The approach of these authors exploits two key institutional features of retail trading. First, most marketable equity orders initiated by retail investors take place off-exchange, and they are either filled from a broker's inventory or routed to wholesalers (Battalio et al. (2016)). Accordingly, we limit our analysis to off-exchange trades, which are designated with the exchange code "D" in TAQ. Second, retail traders typically receive a small price improvement, i.e., a small fraction of a cent, relative to the national best bid or offer (NBBO). Common price improvement amounts include 0.01, 0.1, and 0.2 cents. In contrast, institutional orders tend to be executed at whole or half-cent increments. Thus, we further identify a trade as a retail buy (sell) transaction if it takes place at a price just below (above) a round penny, i.e., trades with fractional penny prices between 0.006 and 0.01 (between 0.00 and 0.004). According to Boehmer et al. (2021), this approach can be used to identify the majority of overall retail trading activity. To measure retail investors' directional trades, we compute the order imbalance of stock i on day d: $OIBVOL_{i,d} = \frac{BVOL_{i,d}-SVOL_{i,d}}{BVOL_{i,d}+SVOL_{i,d}}$, where $BVOL_{i,d}$ and $SVOL_{i,d}$ are the buy and sell volumes of marketable retail orders, respectively, following Boehmer et al. (2021).

Our second proxy for retail trading is based on the number of Robinhood users. Robinhood is an online retail brokerage company that offers commission-free trading on an easy-to-use mobile app. As of June 2021, Robinhood had 22.5 million funded accounts, and more than half of its customers are first-time investors (Darbyshire et al. (2021)). We obtain Robinhood data from the Robintrack website, which provides hourly data on the number of Robinhood users who hold a specific stock.[27] We merge the Robintrack data with CRSP data using the ticker on Robintrack, and identify the last

[27]See more details on the Robintrack website, `https://robintrack.net/`, and in other papers using the same data (e.g., Barber et al. (2021); Welch (2021)). Our analyses using the Robintrack data start from May 2018 due to data availability.

observed user count prior to the close of trading (4 pm ET) for each stock i on each day d (denoted as $user_{i,d}$). We follow Barber et al. (2021) to construct two measures for changes in stock popularity: (1) the daily change in the number of Robinhood users, i.e., $RHNUM_{i,d} = user_{i,d} - user_{i,d-1}$, and (2) the daily percentage change in the number of Robinhood users, i.e., $RHPCT_{i,d} = \frac{user_{i,d}}{user_{i,d-1}} - 1$.

To measure comovements in retail trading activities, we estimate the following daily regressions:

$$TRA_{i,d} = \alpha_0 + \beta_1 DSTRA_{i,d} + \beta_2 MKTTRA_d + \epsilon_{i,d}, \tag{5}$$

where $TRA_{i,d}$ refers to a list of retail trading proxies of stock i on day d, including $OIBVOL_{i,d}$, $RHNUM_{i,d}$, and $RHPCT_{i,d}$, as defined above. $DSTRA_{i,d}$ and $MKTTRA_d$ are the (equal-weighted) retail trading measures of stock i's data-sharing portfolio and the market portfolio, respectively. The standard errors are clustered by calendar day.

The results are reported in Table 8. First, we find a significant comovement in the retail order imbalance between the focal firms and data-sharing firms (Model 1). Specifically, a 1% increase in the data-sharing firms' order imbalance (i.e., *DSOIBVOL*) is associated with a 0.09% increase in the order imbalance of the focal firm after the market average retail trading activities are controlled.

Next, we replace *DSOIBVOL* with (1) $DSOIBVOL^+$, which equals *DSOIBVOL* if $DSOIBVOL > 0$ and 0 otherwise, and (2) $DSOIBVOL^-$, which equals *DSOIBVOL* if $DSOIBVOL < 0$ and 0 otherwise. Prior work suggests that individual investors are net buyers of attention-grabbing stocks and that attention shocks should lead to net buying (e.g., Odean (1999); Barber and Odean (2008)). If the correlated retail trading between the focal firms and data-sharing firms is driven by attention shocks, we expect to see stronger comovement in net buying than in net selling. As shown in Model 2, the focal firm tends to comove more with the data-sharing portfolio when retail investors are net buyers of the data-sharing firms rather than net sellers, as the comovement nearly doubles in this case, i.e., 0.117 vs. 0.059. The difference in the regression coefficients is also statistically significant at the 5% level (F-statistic = 3.84).

Third, our findings remain intact when retail trading is measured by the change in the number of Robinhood users (Models 5-8). For instance, one additional Robinhood user corresponding to the data-sharing firms (i.e., *DSRHNUM*) is associated with a 0.39 increase in the number of Robinhood users corresponding to the focal firm after the market average of Robinhood adoption is controlled

(Model 5).[28] Overall, we show that online data sharing generates attention spillovers within a cookie network, resulting in more joint search for financial information and correlated trading, especially correlated buying among data-sharing firms.

4.3. Data Sharing, Correlated Investor Behavior, and Return Comovement

To further link the return comovement with data sharing and the economic mechanism via correlated EDGAR search and retail trading, we expand our analysis described in Equation (3) to the following monthly Fama and MacBeth (1973) specification:

$$\begin{aligned} ARCORR_{ij,t} = \alpha_0 + \beta_1 NUMTP_{ij,t-1} + \beta_2 NUMTP_{ij,t-1} \times IBCORR_{ij,t-1} \\ + \beta_3 IBCORR_{ij,t-1} + \boldsymbol{\gamma}' \boldsymbol{N}_{ij,t-1} + \epsilon_{ij,t}, \end{aligned} \tag{6}$$

where $IBCORR_{ij,t-1}$ refers to a set of variables indicating the correlation between the investor behavior related to stock i and that related to stock j in month t: $HUMCORR$ is the correlation of daily human search, $ROBCORR$ is the correlation of daily robot search, $OIBVOLCORR$ is the correlation of the daily order imbalance of share volume, $POSOIBVOL$ is the percentage of common retail buys, $NEGOIBVOL$ is the percentage of common retail sells, $RHNUMCORR$ is the correlation of the daily change in the number of Robinhood users, and $RHPCTCORR$ is the correlation of the daily percentage change in the number of Robinhood users. All the other variables are defined as in Equation (3). Appendix A provides a detailed definition of each variable. We also report Newey and West (1987) adjusted t-statistics.

The parameter of interest is β_2. If the higher return comovement between data-sharing firms is indeed due to correlated investor attention and subsequent trading activities, we expect to see stronger comovement when firms also display more correlated EDGAR search and correlated trading from retail investors. The results are tabulated in Table 9. While data sharing affects only the attention of human investors, we nevertheless include robot search as a placebo test in Models 1-2 but expect to see a positive (insignificant) value of β_2 when interacting with human (robot) search.

[28] We do not further consider positive vs. negative changes in the number of Robinhood users, because our sample period coincides with a rapid increase in aggregate Robinhood holdings (Barber et al. (2021); Welch (2021)). Therefore, we rely on the order imbalance measure constructed from TAQ data to provide an accurate estimate of retail buying and selling activities.

Indeed, we find that online data sharing further enhances residual return comovement when the two stocks are jointly searched on EDGAR by human investors. In contrast, the joint robot search does not affect return comovement through data sharing.

Furthermore, we show that the data-sharing firms with more correlated retail trading show greater residual return comovement. The results are robust to various retail trading measures across all the specifications (Models 3-10). Importantly, as shown in Models 5-6, online data sharing enhances residual return comovement only when retail investors are net buyers of both stocks. In contrast, common selling by retail investors does not affect return comovement through data sharing. Collectively, the findings in this section reinforce our previous observation that data sharing facilitates attention spillover, encourages individual investors to acquire financial information and make information-based buys, and therefore leads to return comovement among data-sharing firms.

5. Additional Analyses

5.1. Implications for Trading Strategy

Our previous analysis suggests that online data sharing enhances financial information acquisition and diffusion, leading to return comovement among data-sharing firms. If data sharing causes comovement through attention and information spillover, we can link data sharing to predictable variation in returns. Suppose that stocks i and j share common third-party cookies, and we label stock i as focal stock and stock j as stock i's data-sharing stock. If stock i's price (i.e., own price) declines while data-sharing stock j's price increases, we conjecture that stock i is relatively undervalued and its price should increase once investors learn about the firm, i.e., the positive news spills over from stock j to i. The data-sharing portfolio return provides a reasonable benchmark to evaluate whether the focal stock is undervalued or overvalued, and our trading strategy exploits the lead-lag return predictability induced by information diffusion.

In particular, at the end of day d, we independently sort stocks into quintile portfolios according to their own returns and data-sharing portfolio returns to generate 25 (5×5) portfolios. The low-(high)-own-return and data-sharing-portfolio-return portfolios comprise the bottom (top) quintile of stocks based on the own return and data-sharing portfolio return, respectively. The data-sharing portfolio return is the average return of all stocks with common third-party cookies, weighted

by the number of common third-party cookies. We compute the equal-weighted return on day $d+1$ for each of the 25 portfolios, with the investment strategy of going long (short) the low-(high)-own-return stocks ("LMH") and the investment strategy of going long (short) the high-(low)-data-sharing-portfolio-return stocks ("HML"). "HL − LH" reports returns for the investment strategy of going long the high-data-sharing-portfolio-return and low-own-return stocks and short the low-data-sharing-portfolio-return and high-own-return stocks.

In addition to raw portfolio returns, we follow Antón and Polk (2014) and report risk-adjusted returns from a five-factor model consisting of the FFC four factors (i.e., market, size, book-to-market, and momentum) and the short-term reversal factor (ST_REV, defined as the loser minus winner return premium) (Jegadeesh (1990)). The standard errors in all estimations are corrected for autocorrelation using the Newey and West (1987) method.

Panel A of Table 10 reports the results. Several findings are noteworthy. First, a stock's own return is negatively associated with future performance across all data-sharing-portfolio-return quintiles, suggesting a short-term reversal at a daily frequency. Second, within each own-return quintile, stocks with high data-sharing portfolio returns outperform those with low data-sharing portfolio returns, although the difference is statistically significant only among stocks with high own returns. More important, stocks with low own-return and high data-sharing-portfolio-return yield a daily return (five factor-adjusted return) of 0.17% (0.13%), while stocks with high own-return and low data-sharing-portfolio-return yield a daily return (five factor-adjusted return) of −0.11% (−0.14%). The long-short portfolio yields a daily return of 0.28% (t-statistic = 12.34) and five factor-adjusted returns of 0.27% (t-statistic = 11.88).

Panel B of Table 10 focuses on five factor-adjusted returns and reports similar statistics for alternative holding periods from day $d+1$ to $d+5$ and day $d+1$ to $d+10$. To increase the power of our tests, we construct daily rebalanced portfolios with overlapping holding periods. Specifically, for the strategy with a 5-day holding period, on any given day $d+1$, the strategy holds five portfolios that are formed on days $d-4$ to d. The return on day $d+1$ is an equal-weighted average of the previously initiated five portfolio returns. Focusing on the data-sharing-stock strategy ("HL − LH"), the long-short portfolio yields a daily five factor-adjusted return of 0.06% (t-statistic = 6.22) over the 5-day window and 0.04% (t-statistic = 5.66) over the 10-day window. Consistent with attention and information spillovers among retail investors, return predictability is short lived, as

88% (67%) of the 5-day (10-day) risk-adjusted return is concentrated on the first day.[29] On the other hand, we do not find any subsequent reversal in stock returns, suggesting that the price adjustment of the data-sharing firms is permanent and is not driven by a temporary price impact.

Unreported results show that value-weighted portfolios do not generate significant trading profits; hence, return predictability is more prominent among small stocks that are likely to be held by retail investors (e.g., Kumar (2009); Bali et al. (2011); Han and Kumar (2013)). Overall, we find that online data sharing mitigates the limited attention among retail investors, and targeted attention shocks enhance financial information acquisition via EDGAR search and help incorporate the new information into stock prices, resulting in a permanent price adjustment for data-sharing firms.

5.2. Cash Flow Comovement in the Long-Term

Note that the stock return comovement among the data-sharing firms in a cookie network can also be driven by correlated cash flows. In the above analyses, we focus on the daily frequency to exploit instant attention shocks and minimize the impact of correlated cash flows. In addition, we exploit a plausible exogenous shock, i.e., the enactment of the CCPA, that affects data sharing but not other firm characteristics, showing that the examined return comovement is not entirely due to the endogenous data-sharing decisions made by firms with correlated cash flows. We also control for the potential similarity in firm fundamentals (e.g., comovement in return on equity and the similarity in industry classification and sales growth) in pairwise analyses, which largely eliminates the effect of correlated cash flows. Therefore, our previous results are unlikely to be driven by cash flow comovement.

In this subsection, we seek to provide complementary analyses on whether there is any cash flow comovement among data-sharing firms in the *long-term*. For example, when an individual is booking a flight on American Airlines' website, which allows the Adobe platform (Adobe Audience Manager) to collect browsing data, an advertisement from a hotel company that works with the same platform (i.e., a data-sharing firm of the Adobe cookie network) could pop up in a few minutes or even seconds. Thus, the individual might also purchase from this connected hotel company, leading

[29] Specifically, $0.268/(0.061 \times 5) = 88\%$, where 0.268 (0.061) is the daily five factor-adjusted return of the "HL − LH" strategy with a 1-day (5-day) holding period from Panel A (Panel B).

to a comovement in cash flows between American Airlines and the hotel company.

In the same spirit of return comovement, we estimate the following annual panel regression to measure cash flow comovement:

$$Char_{i,y} = \alpha_0 + \beta_1 DSChar_{i,y} + \boldsymbol{\gamma}' \boldsymbol{C}_{i,y} + \epsilon_{i,y}, \tag{7}$$

where $Char_{i,y}$ is a list of firm characteristics of stock i in year y: Adv is the advertising expenditures scaled by total assets; $SALE$ is the sales scaled by total assets; and RD is the R&D expenses scaled by total assets. $DSChar_{i,y}$ is the (equal-weighted) firm characteristics of stock i's data-sharing portfolio. Vector $\boldsymbol{C}$ stacks all other control variables that might affect the firm's cash flow, including *M2B*, *Tangibility*, *TotalAsset*, *CR*, *Coverage*, *Zscore*, and *Leverage*. Appendix A provides a detailed definition of each variable. We cluster standard errors at the stock level.

The results are reported in Table 11, with Models 1-3 including year fixed effects and Models 4-6 including year and industry fixed effects to absorb the time trend and time-invariant industry characteristics. First, without controlling for industry fixed effects, we find that focal firms tend to comove with other data-sharing firms in relation to advertising expenditures and sales revenue. For instance, a 1% increase in the data-sharing firms' advertising expenditures (sales) is associated with a 0.45% (0.10%) increase for the focal firm in Model 1 (Model 2).

Next, we conduct a placebo test by investigating R&D expenses. We do not expect comovement in the R&D expenses of the data-sharing firms, because online data sharing should affect only the focal firm's cash flows related to advertising and selling (i.e., consumer decisions) but not R&D policy (i.e., manager decisions). However, if data sharing is just a proxy for the fundamental similarities between firms, R&D expenses could be spuriously correlated with data sharing. We find that the R&D expenses of the data-sharing firms are not correlated (Model 3).

Finally, we find that the cash flow comovement is fully absorbed by industry fixed effects (Models 4-6). The economic interpretation of this phenomenon is that time-invariant industry characteristics play an important role in cash flow comovements among data-sharing firms, e.g., consumers' joint purchases are concentrated on products from the same industry. This finding also reinforces our argument based on our previous analyses that data-sharing-induced return comovement is not driven by cash flow comovement when industry fixed effects and other pairwise similarities are

controlled.

6. Conclusion

Firms connected through a common third-party cookie could reach out to the same set of tracked users. Therefore, these widely used third-party cookies can foster a data-sharing network within which firms face common attention shocks. In this paper, we examine the capital market consequences of such common attention shocks due to data sharing within a cookie network. Consistent with the attention spillover effect, we find strong return comovement among data-sharing firms in the same cookie network. The price adjustment is also permanent, suggesting that the cookie network enhances the information diffusion between data-sharing firms rather than imposing a temporary price impact. An identification test using the enactment of the CCPA as an exogenous shock to the effectiveness and intensity of data sharing further supports a causal link. Return comovement among data-sharing firms is also more pronounced for consumer-related industries and for cookies that are more frequently installed. The latter result highlights the importance of the economy of scope in data sharing: more frequently installed cookies allow platform companies to provide more accurate profiling, resulting in better attention capture.

We further test the comovement in information acquisition and retail trading among data-sharing firms, providing direct evidence on the attention spillover effect. Using EDGAR searches as a proxy for investors' attention and information acquisition activities, we show that data sharing increases the joint search for financial information for firms in the same cookie network, and the effect is concentrated on human searches rather than machine downloads. We also find a significant comovement in retail trading among data-sharing firms, and the results are stronger for net buying than net selling. In addition, we link comovement in information acquisition and retail trading to return comovement by showing that online data sharing enhances return comovement when human investors jointly search for two stocks on EDGAR and buy the two stocks together.

Our paper joins a growing literature on data sharing. We provide first-hand empirical evidence on how data sharing can affect the capital market and document a beneficial effect whereby online data sharing alleviates the limited attention of investors and helps incorporate the new information into stock prices.

References

Abel, A. B., J. C. Eberly, and S. Panageas (2013). Optimal inattention to the stock market with information costs and transactions costs. *Econometrica 81*(4), 1455–1481.

Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar (2021). Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics*, forthcoming.

Acquisti, A., C. Taylor, and L. Wagman (2016). The economics of privacy. *Journal of Economic Literature 54*(2), 442–492.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance 23*(4), 589–609.

Antón, M. and C. Polk (2014). Connected stocks. *The Journal of Finance 69*(3), 1099–1127.

Bali, T. G., N. Cakici, and R. F. Whitelaw (2011). Maxing out: Stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics 99*(2), 427–446.

Barber, B. M., X. Huang, T. Odean, and C. Schwarz (2021). Attention-induced trading and returns: Evidence from Robinhood users. *Working Paper*.

Barber, B. M. and D. Loeffler (1993). The "Dartboard" column: Second-hand information and price pressure. *The Journal of Financial and Quantitative Analysis 28*(2), 273–284.

Barber, B. M. and T. Odean (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies 21*(2), 785–818.

Battalio, R., S. A. Corwin, and R. Jennings (2016). Can brokers have it all? on the relation between make-take fees and limit order execution quality. *The Journal of Finance 71*(5), 2193–2238.

Bergemann, D. and A. Bonatti (2015). Selling cookies. *American Economic Journal: Microeconomics 7*(3), 259–294.

Bergemann, D. and A. Bonatti (2019). Markets for information: An introduction. *Annual Review of Economics 11*, 85–107.

Bergemann, D., A. Bonatti, and T. Gan (2021). The economics of social data. *Working Paper*.

Bernard, D., T. Blackburne, and J. Thornock (2020). Information flows among rivals and corporate investment. *Journal of Financial Economics 136*(3), 760–779.

Blankespoor, E., E. deHaan, and I. Marinovic (2020). Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics 70*(2), 101344.

Blankespoor, E., E. deHaan, J. Wertz, and C. Zhu (2019). Why do individual investors disregard accounting information? The roles of information awareness and acquisition costs. *Journal of Accounting Research 57*(1), 53–84.

Boehmer, E., C. M. Jones, X. Zhang, and X. Zhang (2021). Tracking retail investor activity. *The Journal of Finance 76*(5), 2249–2305.

Bozanic, Z., J. L. Hoopes, J. R. Thornock, and B. M. Williams (2017). IRS attention. *Journal of Accounting Research 55*(1), 79–114.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance 52*(1), 57–82.

Chawla, N., Z. Da, J. Xu, and M. Ye (2016). Information diffusion on social media: Does it affect trading, return, and liquidity? *Working Paper*.

Chen, H., S. Chen, and F. Li (2012). Empirical investigation of an equity pairs trading strategy. *Working Paper*.

Chen, X., L. An, Z. Wang, and J. Yu (2021). Attention spillover in asset pricing. *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.

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

Da, Z., J. Engelberg, and P. Gao (2011). In search of attention. *The Journal of Finance 66*(5), 1461–1499.

Da, Z., J. Hua, C.-C. Hung, and L. Peng (2020). Market returns and a tale of two types of attention. *Working Paper*.

Darbyshire, M., M. Badkar, and E. Platt (2021). Robinhood seeks valuation of up to $35bn in IPO. *Financial Times*.

Davies, J. (2017). Know your cookies: A guide to internet ad trackers. *Digiday*.

Davis, J. L., E. F. Fama, and K. R. French (2000). Characteristics, covariances, and average returns: 1929 to 1997. *The Journal of Finance 55*(1), 389–406.

Dechow, P. M., A. Lawrence, and J. P. Ryans (2016). SEC comment letters and insider sales. *The Accounting Review 91*(2), 401–439.

DellaVigna, S. and J. M. Pollet (2009). Investor inattention and Friday earnings announcements. *The Journal of Finance 64*(2), 709–749.

Drake, M. S., J. Jennings, D. T. Roulstone, and J. R. Thornock (2017). The comovement of investor attention. *Management Science 63*(9), 2847–2867.

Drake, M. S., P. J. Quinn, and J. R. Thornock (2017). Who uses financial statements? A demographic analysis of financial statement downloads from EDGAR. *Accounting Horizons 31*(3), 55–68.

Drake, M. S., D. T. Roulstone, and J. R. Thornock (2012). Investor information demand: Evidence from Google searches around earnings announcements. *Journal of Accounting Research 50*(4), 1001–1040.

Drake, M. S., D. T. Roulstone, and J. R. Thornock (2015). The determinants and consequences of information acquisition via EDGAR. *Contemporary Accounting Research 32*(3), 1128–1161.

Drake, M. S., D. T. Roulstone, and J. R. Thornock (2016). The usefulness of historical accounting reports. *Journal of Accounting and Economics 61*(2-3), 448–464.

Easley, D., S. Huang, L. Yang, and Z. Zhong (2018). The economics of data. *Working Paper*.

El-Khatib, R., K. Fogel, and T. Jandik (2015). CEO network centrality and merger performance. *Journal of Financial Economics 116*(2), 349–382.

Engelberg, J., C. Sasseville, and J. Williams (2012). Market madness? The case of Mad Money. *Management Science 58*(2), 351–364.

Englehardt, S. and A. Narayanan (2016). Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA, pp. 1388–1401. Association for Computing Machinery.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Fama, E. F. and J. D. MacBeth (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy 81*(3), 607–636.

Focke, F., S. Ruenzi, and M. Ungeheuer (2020). Advertising, attention, and financial markets. *The Review of Financial Studies 33*(10), 4676–4720.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry 40*(1), 35–41.

Gilbert, T., S. Kogan, L. Lochstoer, and A. Ozyildirim (2012). Investor inattention and the market impact of summary statistics. *Management Science 58*(2), 336–350.

Grullon, G., S. Underwood, and J. P. Weston (2014). Comovement and investment banking networks. *Journal of Financial Economics 113*(1), 73–89.

Hameed, A., R. Morck, J. Shen, and B. Yeung (2015). Information, analysts, and stock return comovement. *The Review of Financial Studies 28*(11), 3153–3187.

Hameed, A. and J. Xie (2019). Preference for dividends and return comovement. *Journal of Financial Economics 132*(1), 103–125.

Han, B. and A. Kumar (2013). Speculative retail trading and asset prices. *Journal of Financial and Quantitative Analysis 48*(2), 377–404.

Hirshleifer, D., S. S. Lim, and S. H. Teoh (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance 64*(5), 2289–2325.

Ho, T. S. Y. and R. Michaely (1988). Information quality and market efficiency. *The Journal of Financial and Quantitative Analysis 23*(1), 53–70.

Hou, K., C. Xue, and L. Zhang (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies 28*(3), 650–705.

Huang, S., Y. Huang, and T.-C. Lin (2019). Attention allocation and return co-movement: Evidence from repeated natural experiments. *Journal of Financial Economics 132*(2), 369–383.

Huberman, G. and T. Regev (2001). Contagious speculation and a cure for cancer: A nonevent that made stock prices soar. *The Journal of Finance 56*(1), 387–396.

Jegadeesh, N. (1990). Evidence of predictable behavior of security returns. *The Journal of Finance 45*(3), 881–898.

Jegadeesh, N. and S. Titman (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance 48*(1), 65–91.

Jiang, L., J. Liu, and B. Yang (2019). Communication and comovement: Evidence from online stock forums. *Financial Management 48*(3), 805–847.

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

Kedia, S. and S. Rajgopal (2011). Do the SEC's enforcement preferences affect corporate misconduct? *Journal of Accounting and Economics 51*(3), 259–278.

Keloharju, M., S. Knüpfer, and J. Linnainmaa (2012). Do investors buy what they know? Product market choices and investment decisions. *The Review of Financial Studies 25*(10), 2921–2958.

Kumar, A. (2009). Who gambles in the stock market? *The Journal of Finance 64*(4), 1889–1933.

Lee, C. M. C., P. Ma, and C. C. Y. Wang (2015). Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics 116*(2), 410–431.

Li, W. and C. Sun (2019). Information acquisition and expected returns: Evidence from EDGAR search traffic. *Working Paper*.

Liang, B. (1999). Price pressure: Evidence from the "Dartboard" column. *The Journal of Business 72*(1), 119–134.

Liaukonytė, J. and A. Žaldokas (2020). Background noise? TV advertising affects real time investor behavior. *Management Science*, forthcoming.

Liu, Z., M. Sockin, and W. Xiong (2021). Data privacy and consumer vulnerability. *Working Paper*.

Lou, D. (2014). Attracting investor attention through advertising. *The Review of Financial Studies 27*(6), 1797–1829.

Madsen, J. and M. Niessner (2019). Is investor attention for sale? The role of advertising in financial markets. *Journal of Accounting Research 57*(3), 763–795.

Mayer, E. J. (2021). Advertising, investor attention, and stock prices: Evidence from a natural experiment. *Financial Management 50*(1), 281–314.

Merton, R. C. (1971). Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory 3*(4), 373–413.

Murgia, M. and M. Harlow (2019). How top health websites are sharing sensitive data with advertisers. *Financial Times 13*.

Muslu, V., M. Rebello, and Y. Xu (2014). Sell-side analyst research and stock comovement. *Journal of Accounting Research 52*(4), 911–954.

Newey, W. K. and K. D. West (1987). A simple positive-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*, 703–708.

Odean, T. (1998). Volume, volatility, price, and profit when all traders are above average. *The Journal of Finance 53*(6), 1887–1934.

Odean, T. (1999). Do investors trade too much? *American Economic Review 89*(5), 1279–1298.

Peng, L. and W. Xiong (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics 80*(3), 563–602.

Ram, A. and M. Murgia (2019). Data brokers: Regulators try to rein in the 'privacy deathstars'. *Financial Times*.

Ramadorai, T., A. Uettwiller, and A. Walther (2019). The market for data privacy. *Working Paper*.

Ryans, J. (2017). Using the EDGAR log file data set. *Working Paper*.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics 50*(3), 665–690.

Tetlock, P. C. (2011). All the news that's fit to reprint: Do investors react to stale information? *The Review of Financial Studies 24*(5), 1481–1512.

Welch, I. (2021). The wisdom of the Robinhood crowd. *The Journal of Finance*, forthcoming.

Figure 1. Example of Cookie Adoption

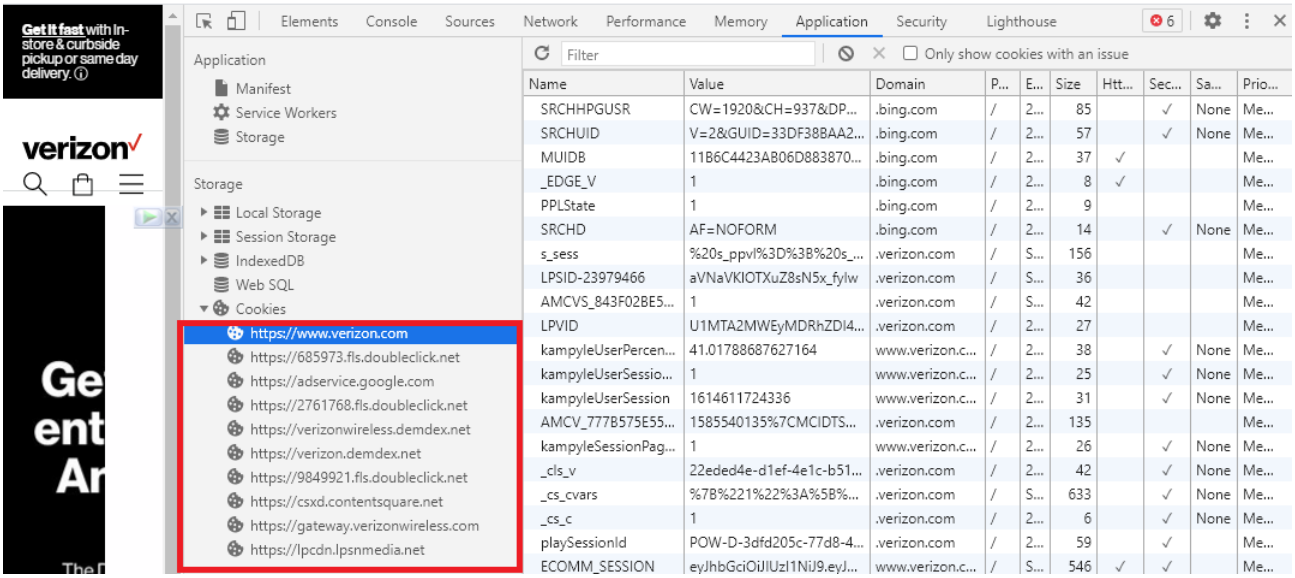

Figure 2. Conceptual Framework of Data Sharing through Cookies

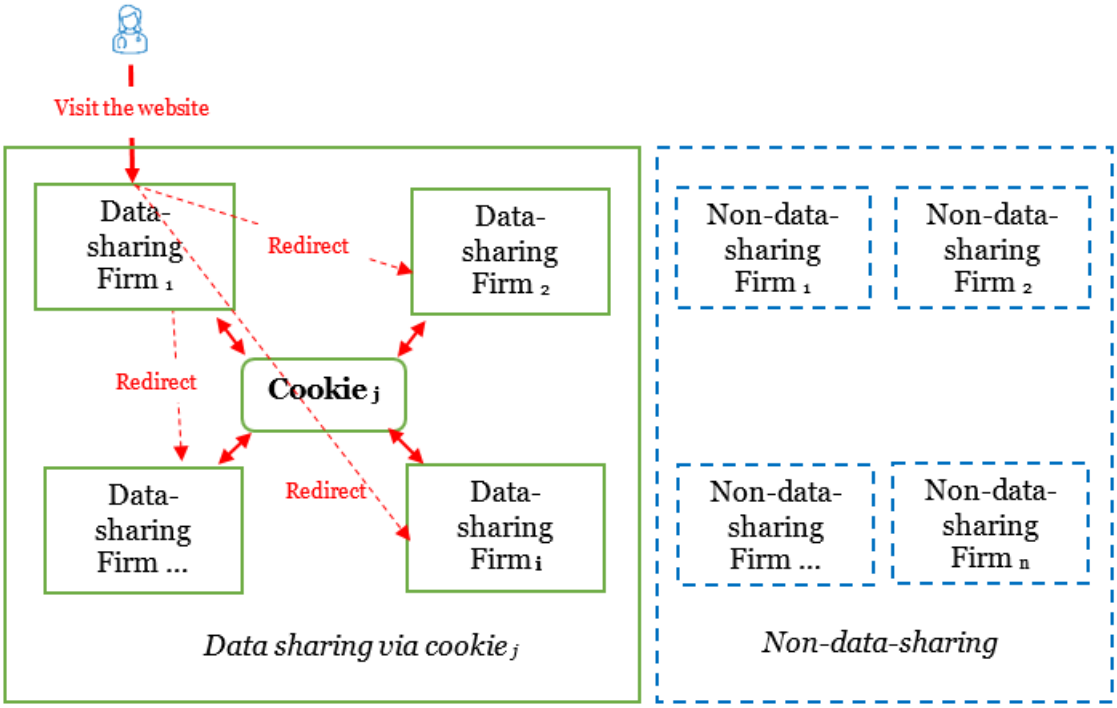

Table 1. List of Top Online Cookie Platforms

This table presents a list of top platforms that own the third-party cookies. For each platform, we report the number of unique industries and the number of unique firms that adopt the platform's third-party cookies, as well as the industry concentration, computed as the Herfindahl-Hirschman Index (HHI) based on the number of adopters in each industry.

Platforms	No. of Firms with Cookies	No. of Industries with Cookies	Industry Concentration
Doubleclick (Google)	959	62	0.073
Facebook	741	56	0.073
LinkedIn	464	49	0.087
Drawbridge Inc	450	49	0.085
Microsoft	333	47	0.068
Twitter	304	44	0.082
TheTradeDesk	274	46	0.073
Adobe	271	41	0.060
AppNexus Inc	220	43	0.075
Yahoo	141	35	0.078
Pardot	109	26	0.099
Rubicon Project	103	33	0.088
Casale Media	100	32	0.075
AddThis	97	33	0.055
OpenX	97	33	0.087
Pubmatic	85	31	0.074
Tapad Inc	84	32	0.059
Exelate	78	27	0.095
Vimeo	78	30	0.076
Lotame	77	23	0.060
Blue Kai	77	31	0.110
Share This	74	26	0.085
MediaMath Inc	74	28	0.097
Eyeota	73	24	0.098
Demandbase	72	22	0.165
Advertising	71	27	0.085
Aggregate Knowledge	69	31	0.056
simpli	65	24	0.129
Quantcast	64	23	0.104
IPONWEB	62	22	0.106

Table 2. Summary Statistics

Panel A presents the summary statistics of firms with at least one third-party cookie in the fiscal year 2019. We report the means, standard deviations, medians, and quantile distributions of a list of firm characteristics. *N* refers to the number of unique firms. Panel B includes all the common stocks listed on NYSE, AMEX, and NASDAQ with available information from CRSP and COMPUSTAT. The stocks are independently sorted according to the number of third-party cookies adopted by the firm and quintile of firm size in 2019. We report the number of stocks in each portfolio. Panel C reports similar statistics for the double-sorted portfolios according to the number of third-party cookies and quintile of market-to-book ratio. Appendix A provides a detailed definition of each variable.

Panel A: Summary Statistics

	N	Mean	p25	p50	p75	S.D.
Ntpcookie	1,348	4.378	1.000	2.000	5.000	5.282
Adv	636	0.025	0.004	0.011	0.033	0.032
SALE	1,344	6.859	5.633	7.123	8.467	2.105
RD	816	0.071	0.004	0.031	0.095	0.098
M2B	1,348	2.143	1.107	1.542	2.665	1.495
Tangibility	1,342	0.262	0.082	0.165	0.403	0.239
TotalAsset	1,348	7.270	5.888	7.451	8.842	2.081
CR	1,246	2.275	1.080	1.619	2.633	1.959
Coverage	1,244	1.881	1.054	1.941	2.669	1.246
Zscore	1,158	2.932	1.092	2.490	4.167	4.289
Leverage	1,348	0.313	0.128	0.301	0.468	0.216

Panel B: Double Sort on the Number of Cookies and Size

	Quintiles of Size					
Ntpcookie	1	2	3	4	5	Total
0	367	379	303	254	168	1,471
1	82	59	86	103	138	468
2	49	39	50	45	55	238
3	25	13	22	34	34	128
4	13	14	23	27	27	104
5	11	15	11	21	24	82
6	2	9	17	15	19	62
7	3	7	10	17	13	50
8	3	5	6	9	17	40
9	3	3	6	7	18	37
$\geq$10	6	21	30	32	50	139
Total	564	564	564	564	563	2,819

Panel C: Double Sort on the Number of Cookies and Market-to-Book Ratio

	Quintiles of Market-to-Book Ratio					
Ntpcookie	1	2	3	4	5	Total
0	315	305	269	287	295	1,471
1	93	82	105	97	91	468
2	49	52	52	52	33	238
3	21	25	24	36	22	128
4	20	16	30	14	24	104
5	17	18	18	14	15	82
6	9	14	9	11	19	62
7	7	7	11	10	15	50
8	6	12	9	8	5	40
9	7	7	12	4	7	37
$\geq$10	20	26	25	31	37	139
Total	564	564	564	564	563	2,819

Table 3. Baseline Results on Return Comovement

This table presents the results of the following daily regressions, as well as their corresponding t-statistics clustered by calendar day:

$$R_{i,d} = \alpha_0 + \beta_1 DSRET_{i,d} + \boldsymbol{\gamma}' \boldsymbol{F}_d + \epsilon_{i,d},$$

where $R_{i,d}$ is the excess return of stock i on day d, and $DSRET_{i,d}$ is the (equal-weighted) excess return of stock i's data-sharing portfolio. Vector $\boldsymbol{F}$ stacks the Fama-French-Carhart (FFC) four factors, including the market factor (MKT), the size factor (SMB), the book-to-market factor (HML), and the momentum factor (MOM). We further replace *DSRET* with *ResidualDSRET* (Model 2), include lagged *DSRET* from day $d-9$ to $d-1$ (Model 3), and include industry fixed effects (Model 4). Panel A reports the regression results over the entire sample period from 2015 to 2019, and Panel B reports similar statistics for 2019 (only the main variables are tabulated for brevity). Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

Panel A: Full Sample (2015–2019)				
	Model 1	Model 2	Model 3	Model 4
DSRET	0.266***		0.255***	0.266***
	(15.91)		(15.82)	(15.91)
ResidualDSRET		0.264***		
		(15.87)		
MKT	0.659***	0.897***	0.670***	0.659***
	(40.89)	(199.57)	(43.21)	(40.86)
SMB	0.486***	0.657***	0.494***	0.486***
	(39.52)	(98.49)	(41.32)	(39.49)
HML	0.108***	0.143***	0.110***	0.108***
	(14.17)	(19.97)	(15.50)	(14.17)
MOM	-0.042***	-0.049***	-0.038***	-0.042***
	(-7.60)	(-8.75)	(-7.26)	(-7.61)
L1.DSRET			0.028***	
			(7.49)	
L2.DSRET			0.004	
			(1.02)	
L3.DSRET			0.007**	
			(2.21)	
L4.DSRET			0.011***	
			(2.77)	
L5_9DSRET			0.001	
			(0.36)	
Constant	0.000**	0.000***	0.000	0.000**
	(2.06)	(2.81)	(1.26)	(2.07)
Industry FE	N	N	N	Y
Obs	1,827,262	1,827,262	1,813,161	1,827,057
R-squared	0.075	0.075	0.075	0.075
Panel B: Recent Year (2019)				
	Model 1	Model 2	Model 3	Model 4
DSRET	0.289***		0.232***	0.290***
	(6.88)		(8.22)	(6.85)
ResidualDSRET		0.289***		
		(6.88)		
L1.DSRET			0.032***	
			(4.41)	
L2.DSRET			0.004	
			(0.53)	
L3.DSRET			0.009	
			(1.16)	
L4.DSRET			0.005	
			(0.64)	
L5_9DSRET			0.003	
			(0.99)	
Constant	0.000	0.000	-0.000	0.000
	(1.22)	(1.41)	(-0.20)	(1.21)
FFC Factors	Y	Y	Y	Y
Industry FE	N	N	N	Y
Obs	363,786	363,786	350,322	363,617
R-squared	0.072	0.072	0.067	0.072

Table 4. Identification Test of Return Comovement and Data Sharing: California Consumer Privacy Act

This table presents the results of the following daily regressions, as well as their corresponding t-statistics clustered by calendar day:

$$R_{i,d} = \alpha_0 + \beta_1 DSRET\_CA_{i,d} + \beta_2 DSRET\_non\text{-}CA_{i,d} + \beta_3 DSRET\_CA_{i,d} \times Post_d + \beta_4 DSRET\_non\text{-}CA_{i,d} \times Post_d + \beta_5 Post_d + \boldsymbol{\gamma}' \boldsymbol{F}_d + \epsilon_{i,d},$$

where $R_{i,d}$ is the excess return of stock i on day d and $DSRET\_CA_{i,d}$ and $DSRET\_non\text{-}CA_{i,d}$ are the excess returns of stock i's data-sharing portfolio with headquarters in and outside California, respectively. $Post_d$ represents several dummy variables: *Post 2Y* equals 1 for two years after the introduction of the California Consumer Privacy Act (CCPA) (i.e., 2018–2019) and 0 otherwise; $Post^{+1}$ equals 1 for one year after the CCPA (i.e., 2018) and 0 otherwise; and $Post^{+2}$ equals 1 for the second year after the CCPA (i.e., 2019) and 0 otherwise. Vector $\boldsymbol{F}$ stacks the Fama-French-Carhart (FFC) four factors. We further include industry fixed effects (Models 4-6). Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
DSRET_CA	0.312***	0.403***	0.406***	0.312***	0.405***	0.407***
	(8.03)	(8.90)	(8.97)	(8.02)	(8.94)	(9.01)
DSRET_non-CA	0.422***	0.418***	0.418***	0.422***	0.418***	0.418***
	(16.91)	(16.64)	(16.77)	(16.93)	(16.65)	(16.79)
DSRET_CA × Post 2Y		-0.165***			-0.167***	
		(-2.62)			(-2.66)	
DSRET_non-CA × Post 2Y		0.010			0.011	
		(0.62)			(0.67)	
DSRET_CA × $Post^{+1}$			-0.082			-0.085
			(-1.03)			(-1.07)
DSRET_non-CA × $Post^{+1}$			-0.007			-0.006
			(-0.33)			(-0.30)
DSRET_CA × $Post^{+2}$			-0.234***			-0.237***
			(-2.92)			(-2.94)
DSRET_non-CA × $Post^{+2}$			0.021			0.022
			(1.07)			(1.11)
Post 2Y		0.000			-0.000	
		(0.08)			(-0.09)	
$Post^{+1}$			-0.000			-0.000
			(-0.39)			(-0.48)
$Post^{+2}$			0.000			0.000
			(0.42)			(0.25)
MKT	0.559***	0.558***	0.557***	0.559***	0.558***	0.557***
	(28.46)	(28.59)	(28.99)	(28.46)	(28.59)	(28.99)
SMB	0.423***	0.420***	0.419***	0.423***	0.419***	0.419***
	(30.10)	(30.03)	(30.22)	(30.09)	(30.03)	(30.22)
HML	0.080***	0.079***	0.079***	0.080***	0.079***	0.079***
	(9.74)	(9.70)	(9.71)	(9.74)	(9.70)	(9.70)
MOM	-0.037***	-0.036***	-0.036***	-0.037***	-0.036***	-0.036***
	(-7.34)	(-7.29)	(-7.31)	(-7.35)	(-7.30)	(-7.31)
Constant	0.000**	0.000**	0.000**	0.000**	0.000**	0.000**
	(2.46)	(1.97)	(1.97)	(2.46)	(2.03)	(2.04)
Industry FE	N	N	N	Y	Y	Y
Obs	1,807,468	1,807,468	1,807,468	1,807,263	1,807,263	1,807,263
R-squared	0.075	0.075	0.075	0.075	0.075	0.075

Table 5. Cross-Sectional Variation in Return Comovement

This table presents the results of the following daily regressions, as well as their corresponding t-statistics clustered by calendar day:

$$R_{i,d} = \alpha_0 + \beta_1 DSRET_{i,d} + \boldsymbol{\gamma}' \boldsymbol{F}_d + \epsilon_{i,d},$$

where $R_{i,d}$ is the excess return of stock i on day d, and $DSRET_{i,d}$ is the (equal-weighted) excess return of stock i's data-sharing portfolio. Vector $\boldsymbol{F}$ stacks the Fama-French-Carhart (FFC) four factors. We further replace *DSRET* with *ResidualDSRET* (Models 2 and 5) and include industry fixed effects (Models 3 and 6). In Panel A, Models 1 to 3 and Models 4 to 6 report results for subsamples of firms in consumer-related industries and other industries, respectively. In Panel B, Models 1 to 3 and Models 4 to 6 report results for subsamples of firms adopting high-frequency cookies and low-frequency cookies, respectively. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

Panel A: Subsamples by Industry Classification						
	Consumer-Related Industries			Other Industries		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
DSRET	0.427***		0.427***	0.180***		0.180***
	(20.29)		(20.35)	(10.19)		(10.16)
ResidualDSRET		0.410***			0.185***	
		(19.73)			(10.40)	
MKT	0.481***	0.864***	0.481***	0.771***	0.932***	0.771***
	(24.50)	(167.46)	(24.56)	(41.84)	(123.15)	(41.77)
SMB	0.331***	0.605***	0.331***	0.598***	0.714***	0.598***
	(20.36)	(68.52)	(20.38)	(37.60)	(58.88)	(37.58)
HML	0.211***	0.268***	0.211***	-0.025*	-0.001	-0.025*
	(18.26)	(24.10)	(18.26)	(-1.96)	(-0.08)	(-1.96)
MOM	0.051***	0.041***	0.052***	-0.144***	-0.149***	-0.145***
	(7.51)	(5.66)	(7.51)	(-15.67)	(-16.28)	(-15.68)
Constant	0.000***	0.000***	0.000***	-0.000	0.000	-0.000
	(2.78)	(3.61)	(2.78)	(-0.03)	(0.26)	(-0.03)
Industry FE	N	N	Y	N	N	Y
Obs	965,834	965,834	965,834	861,428	861,428	861,223
R-squared	0.087	0.086	0.087	0.068	0.068	0.068
Panel B: Subsamples by Cookie Adoption						
	High-Frequency Cookies			Low-Frequency Cookies		
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
DSRET	0.551***		0.551***	0.175***		0.175***
	(24.29)		(24.29)	(17.17)		(17.13)
ResidualDSRET		0.540***			0.174***	
		(23.97)			(17.14)	
MKT	0.396***	0.890***	0.396***	0.767***	0.929***	0.767***
	(18.81)	(304.41)	(18.82)	(74.15)	(205.22)	(74.04)
SMB	0.293***	0.642***	0.293***	0.547***	0.663***	0.547***
	(19.65)	(138.75)	(19.64)	(53.29)	(84.74)	(53.24)
HML	0.067***	0.142***	0.067***	0.124***	0.146***	0.124***
	(11.04)	(27.48)	(11.04)	(15.06)	(18.25)	(15.05)
MOM	-0.024***	-0.036***	-0.024***	-0.034***	-0.040***	-0.034***
	(-6.84)	(-9.56)	(-6.82)	(-5.38)	(-6.26)	(-5.39)
Constant	0.000**	0.000***	0.000**	0.000**	0.000**	0.000**
	(2.25)	(4.29)	(2.25)	(1.98)	(2.50)	(1.98)
Industry FE	N	N	Y	N	N	Y
Obs	1,621,843	1,621,843	1,621,807	1,051,707	1,051,707	1,051,502
R-squared	0.076	0.076	0.076	0.095	0.095	0.095

Table 6. Pairwise Return Comovement and Data Sharing

This table presents the results of the following monthly Fama-MacBeth regressions, as well as their corresponding Newey-West adjusted t-statistics:

$$ARCORR_{ij,t} = \alpha_0 + \beta_1 DS_{ij,t-1} + \boldsymbol{\gamma}' \boldsymbol{N}_{ij,t-1} + \epsilon_{ij,t},$$

where $ARCORR_{ij,t}$ is the correlation of daily Fama-French-Carhart four-factor abnormal returns between stocks i and j in month t. $DS_{ij,t-1}$ refers to a list of data sharing variables for each stock pair: *NUMTP* is the number of common third-party cookies; *LOGNUMTP* is the logarithm of (1+*NUMTP*); *%NUMTP* is defined as *NUMTP* divided by the total number of third-party cookies from the stock pair; and *DNUMTP* is a dummy variable that equals 1 if *NUMTP* > 0 and 0 otherwise. Vector $\boldsymbol{N}$ stacks all other control variables for each stock pair, including *FCAP*, *NUMANA*, *SAMESIZE*, *SAMEBM*, *SAMEMOM*, *NUMSIC*, *SIZE1*, *SIZE2*, and *SIZE1* × *SIZE2*. We also include additional pair controls, including *RETCORR*, *ROECORR*, *VOLCORR*, *DIFFGROWTH*, *DIFFLEV*, *DIFFPRICE*, *DSTATE*, *DINDEX*, and *DLISTING*, as well as style controls, including $SAMESIZE^2$, $SAMESIZE^3$, *BM1*, *BM2*, *BM1* × *BM2*, $SAMEBM^2$, $SAMEBM^3$, *MOM1*, *MOM2*, *MOM1* × *MOM2*, $SAMEMOM^2$, and $SAMEMOM^3$. Appendix A provides a detailed definition of each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
NUMTP	0.185***	0.108***	0.123***	0.125***						
	(11.75)	(8.26)	(8.27)	(8.33)						
LOGNUMTP					0.356***	0.358***				
					(7.78)	(7.83)				
%NUMTP							1.461***	1.416***		
							(7.85)	(7.63)		
DNUMTP									0.277***	0.279***
									(6.78)	(6.88)
FCAP		0.444***	-0.201*	0.514***	-0.209*	0.508***	-0.191*	0.524***	-0.195*	0.521***
		(3.20)	(-1.79)	(5.39)	(-1.87)	(5.30)	(-1.72)	(5.56)	(-1.76)	(5.50)
NUMANA		3.595***	3.536***	3.516***	3.536***	3.516***	3.540***	3.520***	3.540***	3.520***
		(40.95)	(41.21)	(41.50)	(41.21)	(41.50)	(41.24)	(41.53)	(41.21)	(41.50)
SAMESIZE		1.936***	1.264***	6.573***	1.263***	6.566***	1.271***	6.579***	1.262***	6.570***
		(14.33)	(6.92)	(13.70)	(6.89)	(13.68)	(7.01)	(13.72)	(6.91)	(13.69)
SAMEBM		-0.030	-0.244***	2.683***	-0.247***	2.674***	-0.256***	2.680***	-0.251***	2.673***
		(-0.48)	(-3.30)	(7.64)	(-3.33)	(7.63)	(-3.45)	(7.61)	(-3.39)	(7.62)
SAMEMOM		0.892***	0.606***	1.760***	0.603***	1.763***	0.605***	1.757***	0.604***	1.762***
		(10.56)	(6.43)	(4.95)	(6.41)	(4.95)	(6.42)	(4.94)	(6.42)	(4.94)
NUMSIC		1.696***	1.785***	1.773***	1.783***	1.771***	1.791***	1.779***	1.789***	1.777***
		(17.48)	(17.26)	(17.16)	(17.28)	(17.18)	(17.23)	(17.13)	(17.25)	(17.14)
SIZE1		0.267	-0.297	-3.101***	-0.305	-3.108***	-0.254	-3.038***	-0.294	-3.091***
		(0.86)	(-0.83)	(-7.23)	(-0.85)	(-7.25)	(-0.71)	(-7.04)	(-0.82)	(-7.19)
SIZE2		0.164	-0.367	-3.135***	-0.377	-3.142***	-0.333	-3.081***	-0.372	-3.131***
		(0.55)	(-1.09)	(-7.64)	(-1.12)	(-7.65)	(-0.99)	(-7.46)	(-1.10)	(-7.61)
SIZE1 × SIZE2		-0.907	-0.865	3.654***	-0.861	3.658***	-0.877	3.607***	-0.831	3.675***
		(-1.52)	(-1.22)	(4.40)	(-1.22)	(4.40)	(-1.24)	(4.32)	(-1.18)	(4.42)
RETCORR			5.216***	5.379***	5.219***	5.381***	5.205***	5.370***	5.215***	5.378***
			(18.63)	(19.31)	(18.61)	(19.29)	(18.68)	(19.36)	(18.63)	(19.31)
ROECORR			0.261***	0.274***	0.261***	0.274***	0.263***	0.276***	0.263***	0.275***
			(8.20)	(9.67)	(8.21)	(9.67)	(8.19)	(9.65)	(8.20)	(9.66)
VOLCORR			0.531***	0.495***	0.534***	0.498***	0.528***	0.493***	0.533***	0.497***
			(10.39)	(9.85)	(10.48)	(9.94)	(10.34)	(9.81)	(10.45)	(9.91)

Table 6 – Continued

DIFFGROWTH			-0.220***	-0.203***	-0.219***	-0.203***	-0.225***	-0.209***	-0.221***	-0.205***
			(-5.47)	(-5.10)	(-5.48)	(-5.12)	(-5.59)	(-5.21)	(-5.54)	(-5.16)
DIFFLEV			-0.055***	-0.048***	-0.056***	-0.048***	-0.054***	-0.046***	-0.055***	-0.047***
			(-7.91)	(-6.00)	(-8.07)	(-6.15)	(-7.58)	(-5.71)	(-7.85)	(-5.94)
DIFFPRICE			-0.103***	-0.184***	-0.105***	-0.185***	-0.106***	-0.187***	-0.105***	-0.186***
			(-6.18)	(-10.70)	(-6.20)	(-10.71)	(-6.22)	(-10.68)	(-6.19)	(-10.68)
DSTATE			0.036	0.018	0.036	0.018	0.031	0.012	0.033	0.014
			(1.22)	(0.60)	(1.22)	(0.60)	(1.03)	(0.39)	(1.09)	(0.46)
DINDEX			-0.278***	-1.104***	-0.279***	-1.105***	-0.275***	-1.097***	-0.272***	-1.097***
			(-2.95)	(-10.27)	(-2.95)	(-10.27)	(-2.91)	(-10.22)	(-2.89)	(-10.21)
DLISTING			0.391***	0.405***	0.389***	0.403***	0.394***	0.408***	0.391***	0.405***
			(9.23)	(9.59)	(9.20)	(9.56)	(9.27)	(9.64)	(9.22)	(9.58)
SAMESIZE^2				9.601***		9.578***		9.618***		9.596***
				(9.05)		(9.03)		(9.07)		(9.05)
SAMESIZE^3				1.322		1.301		1.365		1.323
				(1.59)		(1.56)		(1.64)		(1.59)
BM1				1.074***		1.063***		1.004***		1.030***
				(4.73)		(4.68)		(4.47)		(4.57)
BM2				1.042***		1.037***		0.973***		1.004***
				(4.39)		(4.37)		(4.16)		(4.26)
BM1 × BM2				-2.854***		-2.832***		-2.715***		-2.768***
				(-5.94)		(-5.91)		(-5.74)		(-5.82)
SAMEBM^2				5.053***		5.035***		5.100***		5.058***
				(5.35)		(5.34)		(5.36)		(5.34)
SAMEBM^3				3.239***		3.214***		3.218***		3.205***
				(4.33)		(4.31)		(4.29)		(4.29)
MOM1				2.587***		2.576***		2.588***		2.586***
				(7.21)		(7.19)		(7.28)		(7.24)
MOM2				2.627***		2.617***		2.628***		2.627***
				(7.35)		(7.33)		(7.44)		(7.39)
MOM1 × MOM2				-4.938***		-4.915***		-4.939***		-4.935***
				(-6.52)		(-6.50)		(-6.58)		(-6.55)
SAMEMOM^2				-2.767***		-2.745***		-2.776***		-2.761***
				(-2.70)		(-2.68)		(-2.71)		(-2.69)
SAMEMOM^3				-3.520***		-3.513***		-3.528***		-3.519***
				(-4.63)		(-4.62)		(-4.64)		(-4.61)
Constant	0.208***	0.479**	0.252	1.280***	0.220	1.253***	0.229	1.279***	0.221	1.261***
	(9.20)	(2.52)	(1.17)	(4.04)	(1.02)	(3.95)	(1.06)	(4.03)	(1.02)	(3.96)
Obs	78,757,999	78,757,999	54,787,691	54,787,691	54,787,691	54,787,691	54,787,691	54,787,691	54,787,691	54,787,691
R-squared	0.000	0.007	0.010	0.011	0.010	0.011	0.010	0.011	0.010	0.011

Table 7. EDGAR Search Comovement and Data Sharing

This table presents the results of the following daily regressions, as well as their corresponding t-statistics clustered by calendar day:

$$HUM_{i,d} = \alpha_0 + \beta_1 DSHUM_{i,d} + \beta_2 DSROB_{i,d} + \beta_3 MKTHUM_d + \beta_4 MKTROB_d + \epsilon_{i,d},$$

where $HUM_{i,d}$ is the number of page views from human readers (i.e., human search) of stock i on day d; $DSHUM_{i,d}$ and $DSROB_{i,d}$ are the (equal-weighted) number of human searches and robot searches of stock i's data-sharing portfolio, respectively; and $MKTHUM_d$ and $MKTROB_d$ are the (equal-weighted) number of human searches and robot searches of the market portfolio, respectively. We further include industry fixed effects (Models 3-4). Appendix A provides a detailed definition of each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	Model 1	Model 2	Model 3	Model 4
DSHUM	0.801***	0.717***	0.731***	0.707***
	(25.79)	(14.67)	(22.32)	(14.51)
DSROB		0.021***		0.006
		(3.10)		(0.86)
MKTHUM	0.209***	0.293***	0.302***	0.315***
	(5.72)	(5.18)	(7.72)	(5.57)
MKTROB		-0.026***		-0.004
		(-2.94)		(-0.46)
Constant	0.430**	0.318	0.611***	-0.134
	(2.51)	(1.26)	(3.17)	(-0.52)
Industry FE	N	N	Y	Y
Obs	1,579,559	1,579,559	1,562,221	1,562,221
R-squared	0.013	0.013	0.207	0.207

Table 8. Comovement in Retail Trading

This table presents the results of the following daily regressions, as well as their corresponding t-statistics clustered by calendar day:

$$TRA_{i,d} = \alpha_0 + \beta_1 DSTRA_{i,d} + \beta_2 MKTTRA_d + \epsilon_{i,d},$$

where $TRA_{i,d}$ refers to a set of retail trading proxies of stock i on day d, including the retail order imbalance of share volume ($OIBVOL$, Models 1-4), the change in the number of Robinhood users ($RHNUM$, Models 5-6), and the percentage change in the number of Robinhood users ($RHPCT$, Models 7-8). $DSTRA_{i,d}$ and $MKTTRA_d$ are the (equal-weighted) retail trading measures of stock i's data-sharing portfolio and the market portfolio, respectively. We further replace $DSOIBVOL$ with $DSOIBVOL^+$ (which equals $DSOIBVOL$ if $DSOIBVOL > 0$ and 0 otherwise) and $DSOIBVOL^-$ (which equals $DSOIBVOL$ if $DSOIBVOL < 0$ and 0 otherwise) (Models 2 and 4) and include industry fixed effects (Models 3-4, 6, and 8). Appendix A provides a detailed definition of each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	OIBVOL				RHNUM		RHPCT	
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
DSOIBVOL	0.086***		0.088***					
	(8.00)		(8.25)					
DSOIBVOL$^+$		0.117***		0.130***				
		(5.79)		(6.05)				
DSOIBVOL$^-$		0.059***		0.053***				
		(3.71)		(3.03)				
MKTOIBVOL	0.967***	0.969***	0.965***	0.968***				
	(37.51)	(37.91)	(37.31)	(37.86)				
DSRHNUM					0.392***	0.372***		
					(5.40)	(5.15)		
DSRHPCT							0.086*	0.084*
							(1.80)	(1.80)
MKTRHNUM					1.168***	1.229***		
					(8.25)	(8.61)		
MKTRHPCT							0.207	0.206
							(1.46)	(1.45)
Constant	-0.011***	-0.011***	-0.010***	-0.011***	-0.854*	-1.045**	0.004***	0.004***
	(-29.71)	(-27.86)	(-29.27)	(-26.51)	(-1.81)	(-2.12)	(4.48)	(4.46)
Industry FE	N	N	Y	Y	N	Y	N	Y
Obs	2,532,805	2,532,805	2,506,883	2,506,883	919,320	907,768	917,642	906,193
R-squared	0.003	0.003	0.004	0.004	0.008	0.025	0.000	0.001

Table 9. Pairwise Return Comovement, Data Sharing, and Investor Behavior

This table presents the results of the following monthly Fama-MacBeth regressions, as well as their corresponding Newey-West adjusted t-statistics:

$$ARCORR_{ij,t} = \alpha_0 + \beta_1 NUMTP_{ij,t-1} + \beta_2 NUMTP_{ij,t-1} \times IBCORR_{ij,t-1} + \beta_3 IBCORR_{ij,t-1} + \boldsymbol{\gamma}' \boldsymbol{N}_{ij,t-1} + \epsilon_{ij,t},$$

where $ARCORR_{ij,t}$ is the correlation of the daily Fama-French-Carhart four-factor abnormal returns between stocks i and j in month t and $NUMTP_{ij,t-1}$ is the number of common third-party cookies. $IBCORR_{ij,t-1}$ refers to a set of variables indicating correlations in investor behavior, including correlations in the daily EDGAR searches by humans (*HUMCORR*) and by robots (*ROBCORR*) (Models 1-2), the correlation in daily retail order imbalances of share volume (*OIBVOLCORR*, Models 3-4), the percentages of common retail buys (*POSOIBVOL*) and retail sells (*NEGOIBVOL*) (Models 5-6), and correlations in the daily change in the number of Robinhood users (*RHNUMCORR*, Models 7-8) and in daily percentage change in the number of Robinhood users (*RHPCTCORR*, Models 9-10). Vector $\boldsymbol{N}$ stacks all other control variables for each stock pair, including *FCAP*, *NUMANA*, *SAMESIZE*, *SAMEBM*, *SAMEMOM*, *NUMSIC*, *SIZE1*, *SIZE2*, and *SIZE1* × *SIZE2*. We also include additional pair controls, including *RETCORR*, *ROECORR*, *VOLCORR*, *DIFFGROWTH*, *DIFFLEV*, *DIFFPRICE*, *DSTATE*, *DINDEX*, and *DLISTING*, as well as style controls, including $SAMESIZE^2$, $SAMESIZE^3$, *BM1*, *BM2*, *BM1* × *BM2*, $SAMEBM^2$, $SAMEBM^3$, *MOM1*, *MOM2*, *MOM1* × *MOM2*, $SAMEMOM^2$, and $SAMEMOM^3$. Appendix A provides a detailed definition of each variable. Only the main variables are tabulated for brevity. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
NUMTP	0.067**	0.078**	0.133***	0.134***	0.121***	0.123***	0.124***	0.125***	0.123***	0.124***
	(2.27)	(2.73)	(8.54)	(8.45)	(6.69)	(6.74)	(5.34)	(5.49)	(5.43)	(5.58)
NUMTP × HUMCORR	0.142***	0.115***								
	(3.22)	(2.81)								
NUMTP × ROBCORR	-0.014	-0.017								
	(-0.57)	(-0.68)								
NUMTP × OIBVOLCORR			0.032**	0.031**						
			(2.39)	(2.30)						
NUMTP × POSOIBVOL					0.102**	0.091**				
					(2.29)	(2.06)				
NUMTP × NEGOIBVOL					-0.054	-0.050				
					(-1.62)	(-1.47)				
NUMTP × RHNUMCORR							0.064***	0.063***		
							(3.44)	(3.45)		
NUMTP × RHPCTCORR									0.066***	0.065***
									(3.51)	(3.53)
HUMCORR	-0.098	-0.054								
	(-1.39)	(-0.77)								
ROBCORR	0.110***	0.107***								
	(2.96)	(2.98)								
OIBVOLCORR			0.009	0.007						
			(0.43)	(0.32)						
POSOIBVOL					-0.123	-0.068				
					(-1.13)	(-0.67)				
NEGOIBVOL					0.305***	0.266***				
					(3.46)	(3.09)				
RHNUMCORR							0.056*	0.058*		
							(1.90)	(1.97)		
RHPCTCORR									0.057*	0.058*
									(1.89)	(1.96)

Table 9 – Continued

FCAP	-0.090	0.465***	-0.181*	0.449***	-0.181*	0.449***	-0.516**	0.410*	-0.517**	0.407*
	(-0.94)	(4.99)	(-1.78)	(4.90)	(-1.78)	(4.91)	(-2.31)	(1.84)	(-2.32)	(1.83)
NUMANA	3.630***	3.608***	3.430***	3.412***	3.430***	3.412***	3.504***	3.484***	3.504***	3.485***
	(43.25)	(42.77)	(41.43)	(41.70)	(41.47)	(41.74)	(19.49)	(19.47)	(19.50)	(19.49)
SAMESIZE	1.005***	6.643***	1.142***	6.256***	1.147***	6.256***	1.176***	5.832***	1.176***	5.821***
	(5.08)	(8.57)	(6.41)	(13.15)	(6.46)	(13.17)	(4.92)	(7.78)	(4.92)	(7.80)
SAMEBM	-0.413***	2.683***	-0.210**	2.842***	-0.212***	2.835***	-0.066	2.691***	-0.067	2.696***
	(-5.68)	(4.33)	(-2.65)	(7.75)	(-2.68)	(7.78)	(-0.40)	(7.11)	(-0.40)	(7.24)
SAMEMOM	0.546***	1.362***	0.654***	2.047***	0.651***	2.048***	0.528**	2.840***	0.532***	2.835***
	(5.41)	(3.08)	(6.81)	(5.44)	(6.78)	(5.45)	(2.87)	(4.28)	(2.91)	(4.29)
NUMSIC	1.742***	1.737***	1.996***	1.984***	1.996***	1.984***	1.946***	1.928***	1.945***	1.927***
	(12.66)	(12.38)	(17.20)	(17.09)	(17.23)	(17.12)	(9.72)	(9.68)	(9.69)	(9.65)
SIZE1	-0.241	-2.802***	-0.377	-3.225***	-0.362	-3.207***	-1.094*	-4.386***	-1.100*	-4.392***
	(-0.86)	(-9.33)	(-1.00)	(-7.26)	(-0.96)	(-7.19)	(-1.86)	(-5.14)	(-1.87)	(-5.15)
SIZE2	-0.231	-2.756***	-0.453	-3.281***	-0.438	-3.263***	-1.191**	-4.472***	-1.204**	-4.483***
	(-0.89)	(-10.63)	(-1.31)	(-7.84)	(-1.26)	(-7.76)	(-2.10)	(-5.31)	(-2.12)	(-5.33)
SIZE1 × SIZE2	-1.168**	2.952***	-0.741	3.863***	-0.750	3.847***	0.969	6.223***	0.988	6.240***
	(-2.41)	(5.38)	(-1.02)	(4.58)	(-1.03)	(4.54)	(0.76)	(3.68)	(0.78)	(3.69)
Constant	-0.211	0.865***	0.248	1.219***	0.185	1.164***	0.787*	2.194***	0.793*	2.200***
	(-1.31)	(2.88)	(1.13)	(3.82)	(0.84)	(3.71)	(1.88)	(3.86)	(1.89)	(3.87)
Pair Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Style Controls	N	Y	N	Y	N	Y	N	Y	N	Y
Obs	26,722,883	26,722,883	50,069,021	50,069,021	50,069,021	50,069,021	15,353,003	15,353,003	15,344,069	15,344,069
R-squared	0.010	0.010	0.011	0.012	0.011	0.012	0.013	0.013	0.013	0.013

Table 10. Stock Returns Sorted by Own Return and Data-Sharing Portfolio Return

At the end of day d, stocks are independently sorted into quintiles according to their own returns and data-sharing portfolio returns to generate 25 (5×5) portfolios. The low- (high)-own-return and data-sharing-portfolio-return portfolios comprise the bottom (top) quintile of stocks based on the own return and data-sharing portfolio return, respectively. The data-sharing portfolio return is the average return of all stocks with common third-party cookies, weighted by the number of common third-party cookies. Panel A reports the equal-weighted return on day $d+1$ for each of the 25 portfolios, the investment strategy of going long (short) the low- (high)-own-return stocks ("LMH"), and the investment strategy of going long (short) the high- (low)-data-sharing-portfolio-return stocks ("HML"). "HL − LH" reports returns for the investment strategy of going long the high-data-sharing-portfolio-return and low-own-return stocks and short the low-data-sharing-portfolio-return and high-own-return stocks. Portfolio returns are further adjusted by a five-factor model including the Fama-French-Carhart four factors and the short-term reversal factor. Panel B focuses on five factor-adjusted returns and reports similar statistics for alternative holding periods from day $d+1$ to $d+5$ and day $d+1$ to $d+10$. Newey-West adjusted t-statistics are shown in parentheses. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

Panel A: Returns to Investment Strategies Sorted by Own Return and Data-Sharing Portfolio Return (1-Day)

	Return							Five factor-adjusted Return						
Own Return	Data-Sharing Portfolio Return							Data-Sharing Portfolio Return						
	Low	2	3	4	High	HML		Low	2	3	4	High	HML	
Low	0.140***	0.152***	0.155***	0.186***	0.166***	0.027		0.098***	0.111***	0.115***	0.144***	0.125***	0.027	
	(3.89)	(4.35)	(4.76)	(5.57)	(4.77)	(1.44)		(5.48)	(6.57)	(7.56)	(9.19)	(7.32)	(1.45)	
2	0.040	0.061**	0.059**	0.057**	0.042	0.002		0.004	0.023**	0.021***	0.020**	0.005	0.001	
	(1.52)	(2.33)	(2.27)	(2.12)	(1.52)	(0.17)		(0.40)	(2.53)	(2.64)	(2.22)	(0.52)	(0.13)	
3	0.035	0.043*	0.031	0.040	0.049**	0.014		0.002	0.008	-0.004	0.005	0.015**	0.013	
	(1.46)	(1.73)	(1.24)	(1.58)	(1.97)	(1.29)		(0.27)	(1.12)	(-0.56)	(0.56)	(2.06)	(1.13)	
4	0.018	0.024	0.028	0.033	0.026	0.008		-0.015*	-0.009	-0.004	0.000	-0.006	0.010	
	(0.68)	(0.94)	(1.09)	(1.30)	(1.00)	(0.84)		(-1.91)	(-1.05)	(-0.54)	(0.03)	(-0.66)	(0.95)	
High	-0.114***	-0.095***	-0.070**	-0.069**	-0.057	0.057***		-0.144***	-0.126***	-0.100***	-0.099***	-0.084***	0.059***	
	(-3.09)	(-2.96)	(-2.15)	(-2.05)	(-1.58)	(3.12)		(-8.18)	(-7.68)	(-7.22)	(-6.31)	(-4.67)	(3.22)	
LMH	0.253***	0.247***	0.226***	0.255***	0.223***	**0.280*** **	HL − LH	0.242***	0.237***	0.215***	0.243***	0.209***	**0.268*** **	HL − LH
	(10.57)	(10.65)	(10.82)	(11.39)	(8.58)	**(12.34)**		(10.10)	(10.36)	(10.23)	(10.91)	(8.33)	**(11.88)**	

Panel B: Five Factor-adjusted Returns to Investment Strategies Sorted by Own Return and Data-Sharing Portfolio Return (5-Day and 10-Day)

	5-Day							10-Day						
Own Return	Data-Sharing Portfolio Return							Data-Sharing Portfolio Return						
	Low	2	3	4	High	HML		Low	2	3	4	High	HML	
Low	0.018	0.029***	0.030***	0.039***	0.029**	0.011		0.015	0.020**	0.018**	0.019**	0.019*	0.004	
	(1.44)	(3.02)	(3.12)	(4.01)	(2.45)	(1.33)		(1.23)	(2.23)	(2.16)	(2.16)	(1.72)	(0.72)	
2	0.000	0.011**	0.010**	0.013***	0.003	0.002		0.004	0.011**	0.008**	0.007*	0.003	-0.001	
	(0.04)	(2.22)	(1.97)	(2.66)	(0.43)	(0.47)		(0.71)	(2.55)	(2.02)	(1.78)	(0.53)	(-0.21)	
3	0.007	0.005	0.003	0.010**	0.007	-0.000		0.008*	0.010***	0.007*	0.009**	0.005	-0.003	
	(1.46)	(1.26)	(0.81)	(2.20)	(1.39)	(-0.09)		(1.87)	(2.71)	(1.83)	(2.32)	(1.04)	(-1.00)	
4	-0.005	0.007	0.003	0.008	0.006	0.010**		0.002	0.006	0.003	0.004	0.006	0.004	
	(-0.88)	(1.41)	(0.69)	(1.64)	(1.11)	(2.18)		(0.42)	(1.57)	(0.75)	(1.07)	(1.32)	(1.03)	
High	-0.032***	-0.028***	-0.021**	-0.013	-0.019*	0.013*		-0.021**	-0.015*	-0.008	-0.006	-0.013	0.008	
	(-2.91)	(-3.10)	(-2.43)	(-1.53)	(-1.85)	(1.70)		(-2.12)	(-1.83)	(-1.11)	(-0.85)	(-1.39)	(1.47)	
LMH	0.050***	0.058***	0.051***	0.053***	0.048***	**0.061*** **	HL − LH	0.036***	0.034***	0.026***	0.025***	0.032***	**0.040*** **	HL − LH
	(4.80)	(6.18)	(5.18)	(5.33)	(4.57)	**(6.22)**		(5.02)	(5.51)	(4.03)	(4.20)	(4.29)	**(5.66)**	

Table 11. Cash Flow Comovement in the Long Term

This table presents the results of the following annual panel regressions with year fixed effects and the corresponding t-statistics with standard errors clustered at the stock level:

$$Char_{i,y} = \alpha_0 + \beta_1 DSChar_{i,y} + \boldsymbol{\gamma}' \boldsymbol{C}_{i,y} + \epsilon_{i,y},$$

where $Char_{i,y}$ is a list of firm characteristics of stock i in year y: *Adv* is the advertising expenditures scaled by total assets (Models 1 and 4); *SALE* is the sales scaled by total assets (Models 2 and 5); and *RD* is the R&D expenses scaled by total assets (Models 3 and 6). $DSChar_{i,y}$ represents the (equal-weighted) firm characteristics of stock i's data-sharing portfolio. Vector $\boldsymbol{C}$ stacks all other control variables, including *M2B*, *Tangibility*, *TotalAsset*, *CR*, *Coverage*, *Zscore*, and *Leverage*. We further include industry fixed effects (Models 4-6). Appendix A provides a detailed definition of each variable. Numbers with "*", "**", and "***" are significant at the 10%, 5%, and 1% levels, respectively.

	Advertising	Sales	R&D	Advertising	Sales	R&D
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
DSAdv	0.454*			0.133		
	(1.84)			(0.53)		
DSSALE		0.095**			0.033	
		(2.22)			(0.85)	
DSRD			0.162			-0.021
			(0.83)			(-0.11)
M2B	0.003***	0.001	0.009***	0.003***	0.001	0.009***
	(4.25)	(0.65)	(18.72)	(4.09)	(1.23)	(17.37)
Tangibility	0.011	-0.763***	-0.091***	0.007	-0.376***	-0.039
	(1.23)	(-7.52)	(-5.62)	(0.60)	(-2.67)	(-1.57)
TotalAsset	-0.003**	0.938***	-0.012***	-0.002**	0.945***	-0.012***
	(-2.46)	(83.38)	(-5.17)	(-2.06)	(85.14)	(-4.75)
CR	-0.001	-0.080***	-0.000	-0.001	-0.066***	-0.002
	(-1.60)	(-6.40)	(-0.15)	(-1.21)	(-6.06)	(-1.50)
Coverage	-0.002	0.247***	-0.022***	-0.004*	0.184***	-0.018***
	(-0.95)	(14.55)	(-9.79)	(-1.89)	(12.99)	(-8.21)
Zscore	0.000	0.000	-0.001	0.000	0.000	-0.001
	(0.11)	(0.25)	(-1.22)	(0.11)	(0.03)	(-1.21)
Leverage	-0.027**	0.446***	-0.089***	-0.026**	0.424***	-0.097***
	(-2.40)	(3.71)	(-3.41)	(-2.30)	(3.73)	(-3.52)
Constant	0.040***	-0.860***	0.204***	0.017	-0.821***	0.144***
	(2.66)	(-3.07)	(7.79)	(0.92)	(-2.75)	(6.31)
Year FE	Y	Y	Y	Y	Y	Y
Industry FE	N	N	N	Y	Y	Y
Obs	2,838	7,002	4,216	2,838	7,002	4,216
R-squared	0.446	0.879	0.367	0.484	0.908	0.392

Appendices

A. Variable Definitions

Variables	Definitions
Panel A. Stock and Data-Sharing Portfolio Characteristics	
Ntpcookie	The number of unique third-party cookie platforms adopted by the firm.
DSRET	The daily return of the data-sharing portfolio.
HUM	The daily EDGAR page views corresponding to human readers.
DSHUM	The daily EDGAR page views corresponding to human readers of the data-sharing portfolio.
ROB	The daily EDGAR page views corresponding to automated machine downloads (i.e., robots).
DSROB	The daily EDGAR page views corresponding to robots of the data-sharing portfolio.
OIBVOL	The retail order imbalance for share volume for stock i on day d is computed as follows: $OIBVOL_{i,d} = \frac{BVOL_{i,d}-SVOL_{i,d}}{BVOL_{i,d}+SVOL_{i,d}}$, where $BVOL_{i,d}$ and $SVOL_{i,d}$ are the buy and sell volume from marketable retail orders, respectively, following Boehmer et al. (2021).
DSOIBVOL	The daily retail order imbalance of the share volume of the data-sharing portfolio.
DSOIBVOL^{+}	A variable that equals *DSOIBVOL* if $DSOIBVOL > 0$ and 0 otherwise.
DSOIBVOL^{-}	A variable that equals *DSOIBVOL* if $DSOIBVOL < 0$ and 0 otherwise.
MKTOIBVOL	The daily retail order imbalance of the share volume of the market portfolio.
RHNUM	The change in the number of Robinhood users corresponding to stock i on day d is computed as follows: $RHNUM_{i,d} = user_{i,d} - user_{i,d-1}$, where $user_{i,d}$ is the last observed Robinhood user count prior to the close of trading (4 pm ET), following Barber et al. (2021).
DSRHNUM	The daily change in the number of Robinhood users corresponding to the data-sharing portfolio.
MKTRHNUM	The daily change in the number of Robinhood users corresponding to the market portfolio.
RHPCT	The percentage change in the number of Robinhood users holding stock i on day d is computed as follows: $RHPCT_{i,d} = \frac{user_{i,d}}{user_{i,d-1}} - 1$, where $user_{i,d}$ is defined as in *RHNUM*, following Barber et al. (2021).
DSRHPCT	The daily percentage change in the number of Robinhood users corresponding to the data-sharing portfolio.
MKTRHPCT	The daily percentage change in the number of Robinhood users corresponding to the market portfolio.
Adv	Advertising expenditures scaled by total assets.
DSAdv	The *Adv* (advertising expenditures scaled by total assets) of the data-sharing portfolio.
SALE	Sales scaled by total assets.
DSSALE	The *SALE* (sales scaled by total assets) of the data-sharing portfolio.
RD	R&D expenses scaled by total assets.
DSRD	The *RD* (R&D expenses scaled by total assets) of the data-sharing portfolio.
M2B	The market value of equity plus the book value of debt scaled by total assets.
Tangibility	Net property, plant and equipment scaled by total assets.
TotalAsset	The logarithm of total assets.
CR	Current assets scaled by current liabilities.
Coverage	EBIT scaled by interest payments on debt.
Zscore	1.2 × working capital / total assets + 1.4 × retained earnings / total assets + 3.3 × EBIT / total assets + 0.6 × market value of equity / total liabilities + 1.0 × sales / total assets, following Altman (1968).
Leverage	Long term debt plus debt in current liabilities scaled by the total assets.
Panel B. Stock Pair Characteristics	
ARCORR	The correlation of a stock pair's daily four-factor abnormal return in a month, in percentage. The abnormal return is computed as the realized stock return minus the product of a stock's four-factor betas and the realized four-factor returns. The four-factor model consists of Fama and French (1993) and Carhart (1997) factors (market, size, book-to-market, and momentum). The betas of the fund are estimated as the exposures of the stock to the relevant risk factors using daily data in a given month.
HUMCORR	The correlation of a stock pair's daily EDGAR search by humans (i.e., *HUM* defined above) in a month.
ROBCORR	The correlation of a stock pair's daily EDGAR search by robots (i.e., *ROB* defined above) in a month.
OIBVOLCORR	The correlation of a stock pair's daily retail order imbalance of share volume (i.e., *OIBVOL* defined above) in a month.
POSOIBVOL	The percentage of common retail buys between a stock pair, defined as the number of days that both stocks have positive retail order imbalances of share volume (i.e., *OIBVOL* defined above) divided by the number of trading days in a month.
NEGOIBVOL	The percentage of common retail sells between a stock pair, defined as the number of days that both stocks have negative retail order imbalances of share volume (i.e., *OIBVOL* defined above) divided by the number of trading days in a month.
RHNUMCORR	The correlation of a stock pair's daily change in the number of Robinhood users (i.e., *RHNUM* defined above) during a month.

Appendix A – Continued

RHPCTCORR	The correlation of a stock pair's daily percentage change in the number of Robinhood users (i.e., *RHPCT* defined above) during a month.
NUMTP	The number of common third-party cookies shared between a stock pair.
LOGNUMTP	The logarithm of one plus the number of common third-party cookies shared between a stock pair.
%NUMTP	The percentage number of common third-party cookies shared between a stock pair in a given month t is computed as follows: $\%NUMTP_{ij,t} = 2 \times NUMTP_{ij,t} / (NUMTP_{i,t} + NUMTP_{j,t})$, where $NUMTP_{ij,t}$ is the number of common third-party cookies covered by both stocks i and j in month t and $NUMTP_{i,t}$ ($NUMTP_{j,t}$) is the number of third-party cookies covered by stock i (j).
DNUMTP	A dummy variable that equals 1 if there exist common third-party cookies between a stock pair and 0 otherwise.
FCAP	The common ownership in the two stocks in quarter q is computed as follows: $FCAP_{ij,t} = \sum_{f=1}^{F} \left(S_{i,q}^{f} P_{i,q} + S_{j,q}^{f} P_{j,q}\right) / (S_{i,q} P_{i,q} + S_{j,q} P_{j,q})$, where $S_{i,q}^{f}$ ($S_{j,q}^{f}$) is the number of shares of stock i (j) held by a common fund f in quarter q, $P_{i,q}$ ($P_{j,q}$) is the price of stock i (j), and $S_{i,q}$ ($S_{j,q}$) is the number of shares outstanding of stock i (j). F indicates the total number of common funds that hold both stocks i and j in their portfolios.
NUMANA	The number of analysts that issued at least one annual earnings forecast for both stocks in the last year.
SIZE	The market capitalization of a stock computed as the number of common shares outstanding times share price. In addition, *SIZE1* and *SIZE2* are defined as the normalized rank-transform of the percentile *SIZE* of the two stocks in a stock pair.
SAMESIZE	The negative of the absolute difference in the two stocks' percentile ranking of *SIZE*. In addition, $SAMESIZE^2$ and $SAMESIZE^3$ are the square and cube of *SAMESIZE*, respectively.
BM	The book value of equity divided by market capitalization at fiscal year-end. The book value of equity is computed as the stockholders' equity (COMPUSTAT annual item SEQ), plus deferred taxes (item TXDB) and investment tax credit (item ITCB), minus the book value of the preferred stock. Depending on availability, we use the redemption value (item PSTKRV), liquidation value (item PSTKL), or carrying value (item PSTK) to estimate the book value of the preferred stock, following Fama and French (1993), and Davis et al. (2000). In addition, *BM1* and *BM2* are defined as the normalized rank-transform of the percentile *BM* of the two stocks in a stock pair.
SAMEBM	The negative of the absolute difference in the two stocks' percentile ranking of *BM*. In addition, $SAMEBM^2$ and $SAMEBM^3$ are the square and cube of *SAMEBM*, respectively.
MOM	The past return in a given month t is computed as the cumulative 12-month return from month $t-12$ to month $t-1$, following Jegadeesh and Titman (1993). In addition, *MOM1* and *MOM2* are defined as the normalized rank-transform of the percentile *MOM* of the two stocks in a stock pair.
SAMEMOM	The negative of the absolute difference in the two stocks' percentile ranking of *MOM*. In addition, $SAMEMOM^2$ and $SAMEMOM^3$ are the square and cube of *SAMEMOM*, respectively.
NUMSIC	The number of consecutive SIC digits, beginning with the first digit, that is equal for a stock pair.
RETCORR	The correlation of a stock pair's monthly return in the last five years.
ROECORR	The correlation of stock pair's quarterly return on equity (ROE) in the last five years. ROE is a given quarter q is computed as follows: $ROE_{i,q} = INCOME_{i,q} / EQUITY_{i,q-1}$, where $INCOME_{i,q}$ is the income before extraordinary items (COMPUSTAT quarterly item IBQ) of stock i in quarter q and $EQUITY_{i,q-1}$ is the shareholders' equity. Depending on availability, we use stockholders' equity (item SEQQ), common equity (item CEQQ) plus redemption value (item PSTKRQ), common equity (item CEQQ) plus the carrying value of the preferred stock (item PSTKQ), or total assets (item ATQ) minus total liabilities (item LTQ) in that order as shareholders' equity, following Hou et al. (2015).
VOLCORR	The correlation of a stock pair's monthly abnormal trading volume in the last five years. The abnormal trading volume is computed as the residual from a regression of monthly trading volume on annual trend and monthly dummies with data from the last three years, following Chen et al. (2012).
DIFFGROWTH	The absolute difference in the two stocks' five-year log sales growth rate in year y is computed as follows: $DIFFGROWTH_{ij,y} = \lvert \log(Sales_{i,y}/Sales_{i,y-5}) - \log(Sales_{j,y}/Sales_{j,y-5}) \rvert$, where $Sales_{i,y}$ ($Sales_{j,y}$) is the sales (COMPUSTAT annual item SALE) of stock i (j) in year y.
DIFFLEV	The absolute difference in the two stocks' financial leverage ratio, defined as long-term debt (COMPUSTAT quarterly item LTQ) divided by total assets (item ATQ).
DIFFPRICE	The absolute difference in the two stocks' log share price.
DSTATE	A dummy variable that equals 1 if the two firms are located in the same state and 0 otherwise.
DINDEX	A dummy variable that equals 1 if the two stocks belong to the S&P 500 index and 0 otherwise.
DLISTING	A dummy variable that equals 1 if the two stocks are listed on the same stock exchange and 0 otherwise.