Open-Source Activities and Stock Returns: Evidence from GitHub Engagement

Cagri Berk Onuk* Kuntara Pukthuanthong† Yunfei Zhao‡
Wenzhou-Kean University *University of Missouri* *Wenzhou-Kean University*

February 24, 2026

Abstract

We examine whether GitHub engagement predicts stock returns. Using data linking S&P 1500 firms to GitHub activity (2010–2022), we construct a market-equity scaled PCA measure of engagement based on GitHub stars, forks, watchers, and repositories. Portfolios sorted on this composite measure yield significant abnormal returns: the high-minus-low spread is 1.45% monthly, compounding to 52.48% over five years, with robust Fama-French and q-factor alphas. Predictability strengthens for hard-to-fake metrics and survives fake-star cleansing. GitHub engagement predicts returns at intermediate horizons (peaking at a five-month lag) but not at short horizons, contrasting with traditional innovation proxies that predict at immediate lags, consistent with a production-side signal that materializes through integration, deployment, and adoption cycles. A decomposition shows GitHub stars reflect future profitability and intangible value. These findings suggest open-source engagement signals organization capital that investors underweight, contributing to literatures on alternative data and intangible assets.

Keywords: GitHub, Open-Source Platforms, Return Predictability, Intangible Capital, Alternative Data, Innovation, Asset Pricing
JEL Classification: G12, O32, D83, L86

*College of Business and Public Management, Wenzhou-Kean University, Wenzhou, China. conuk@kean.edu.
†Trulaske College of Business, University of Missouri, Columbia, MO. pukthuanthongk@missouri.edu.
‡College of Business and Public Management, Wenzhou-Kean University, Wenzhou, China. yunzhao@wku.edu.cn.

1 Introduction

Financial markets increasingly rely on alternative data to infer firm fundamentals, yet not all signals are equally incorporated into prices. We study a class of signals that are publicly observable but costly to process, which theory predicts will be underweighted by investors and thus generate predictable returns (Grossman and Stiglitz, 1980; Hong and Stein, 1999).

Firms increasingly rely on open-source platforms not only to build and deploy software but also to signal their technological capabilities and participate in broader innovation ecosystems. Markets also increasingly rely on digital footprints to infer firms' technological capabilities and intangible assets. Among these footprints, corporate engagement with GitHub—the leading open-source platform—offers a real-time, community-validated view of software activity that is largely absent from standard financial data pipelines. We ask a predictive question: *Do open-source digital footprints predict cross-sectional variation in future stock returns and fundamentals?*

Our setting is open-source software engagement on GitHub, the dominant platform for collaborative coding. GitHub activity is public—anyone can view repositories, stars, forks, and contributions—but it is outside standard investor data pipelines and requires substantial technical and data-engineering effort to interpret. Unlike structured accounting data, GitHub signals are high-dimensional, noisy, and embedded in a developer ecosystem that most investors lack the expertise to parse. These frictions create a natural laboratory for testing whether markets underreact to complex, off-pipeline information.

Firms' decisions to share code on GitHub are not incidental, but reflect a deliberate trade-off between proprietary costs and strategic benefits. Open-source participation allows firms to attract and screen skilled developers, build reputation within technical communities, accelerate innovation through external contributions, and increase adoption of complementary technologies (West and Gallagher, 2006; Lerner and Tirole, 2002). By making portions of their codebase publicly visible, firms can credibly signal technological sophistication, openness, and commitment to modern development practices—attributes that are difficult to convey through traditional

disclosures. Recent evidence from labor markets underscores this signaling role: a 2016 GitHub policy change that increased the visibility of individual contributions led employees with higher open-source activity to experience significantly greater mobility toward large firms, suggesting that GitHub engagement conveys valuable information about underlying organization capital and technical capability (Gupta et al., 2024). While such disclosure entails risks of imitation or intellectual property leakage, these costs are often outweighed by the benefits of ecosystem participation, talent acquisition, and reputational capital, particularly for firms operating in software- and data-intensive industries (Lerner and Tirole, 2005).

While GitHub is widely used by individual developers and organizations alike, the economic and strategic implications of corporate GitHub activity remain underexplored in the finance and innovation literatures. Unlike traditional innovation metrics such as R&D expenditures or patent counts, which are often released with delays and overlook early-stage development, collaboration, and open-source contributions (Reeb and Zhao, 2020), GitHub activity is public, contemporaneous, and observable in real time (Maxfield Brown et al., 2024). Firms that actively share code and contribute to open-source projects may be perceived as more technologically sophisticated, open to collaboration, and forward-looking. As such, GitHub engagement may serve as a valuable lens through which to understand both firm behavior and market expectations.

We examine whether open-source engagement, as a digital footprint of technical activity, predicts cross-sectional variation in future stock returns and fundamentals. Using a hand-curated mapping of S&P 1500 firms to GitHub organizations (2010–2022), we show that firms with higher GitHub engagement earn large and persistent return spreads that remain after controlling for standard factors and traditional innovation proxies (patents, R&D, trademarks). To capture overall engagement, we construct a composite GitHub activity measure using a market-equity-scaled principal component analysis (PCA) of multiple platform metrics, i.e., the number of stars, forks, watchers, and repositories. Scaling by market equity mitigates the mechanical correlation between GitHub activity and firm size, as larger firms tend to operate more repositories and attract greater attention from the developer community. This adjustment

allows us to isolate variation in engagement that is not simply driven by firm scale. Sorting firms monthly into terciles based on this engagement factor, the high-minus-low (H-L) portfolio delivers an excess return of 1.45% per month ($t = 2.91$), with FF5 and q-factor alphas of 1.55% and 1.72% per month, respectively (Table 4). The spread compounds to 52.48% with a t-statistic of 9.18 over 60 months, with positive FF5 and q-alphas across horizons (Table 5). The persistence of returns over such long horizons is consistent with a systematic risk interpretation, as pure mispricing would typically be arbitraged away more quickly. However, it is also consistent with a network-externality mechanism in which early GitHub engagement initiates ecosystem feedback loops that amplify firm value over time. Predictability strengthens for hard-to-fake metrics (forks, pull requests, contributors, issues, and commits) and after cleansing fake stars. This pattern is difficult to reconcile with a pure attention or manipulation story. If GitHub engagement merely captured noisy hype or artificially inflated popularity, return predictability would weaken once suspect activity is removed and when focusing on engagement measures that are costly to fake. Instead, the strengthening of results under these more stringent specifications indicates that GitHub metrics contain economically meaningful information rather than noise. Our findings therefore reflect not mechanical popularity effects, but a real innovation signal that survives rigorous manipulation checks and alternative measurement approaches. GitHub engagement also forecasts future ROE and intangible value added, consistent with an organization-capital channel.

We conjecture that firms' open-source sharing on GitHub provides a forward-looking signal of stock returns because it reveals real-time information about innovation effort, organization capital, and ecosystem engagement that precedes observable financial outcomes. Although these signals are publicly available, they are technically complex and costly to process, leading to gradual incorporation into prices (Hong and Stein, 1999). At the same time, GitHub engagement may proxy for exposure to systematic innovation risk, implying higher expected returns as compensation. Together, these channels make GitHub activity informative about both future fundamentals and return dynamics.

GitHub stars serve as a form of community-based endorsement and recognition, potentially reflecting both the utility and perceived quality of a firm's technological output (Borges and Valente, 2018). At the same time, the computer science literature has recently highlighted a number of caveats when interpreting the number of stars a GitHub repository receives. For example, Koch et al. (2024) point out that the correlation between the number of stars and downloads can drop as low as 0.14 for repositories in certain programming languages. This implies that GitHub stars can sometimes proxy for attention rather than code utility. Furthermore, He et al. (2024) demonstrate the growing fake star issue on GitHub. In particular, the star count is sometimes at risk of being artificially inflated. For instance, malware repositories disguised as game cheats or cryptocurrency bots are promoted via fake star campaigns. In such campaigns, multiple bots star a repository in lockstep patterns. According to He et al. (2024), fake stars reached a peak in July 2024, when 15.84% of repositories with greater than 50 stars were targets of fake star campaigns. Notwithstanding the fake-star and attention-versus-innovation issues, the literature also finds that three out of four developers still consider the number of stars before using or contributing to a GitHub project (Borges and Valente, 2018). The same study also finds that practitioners view stars as the most useful measure of popularity on GitHub.

Ultimately, we measure GitHub engagement using *stars*, *forks*, *watchers*, and other *repository activity*. We develop and test a limited-attention framework in which these GitHub metrics are economically meaningful yet costly to process for most investors, yielding gradual information diffusion and return predictability (Hong and Stein, 1999). The economics are consistent with GitHub activity proxying for organization capital and ecosystem value that are observable in real time but imperfectly priced (Eisfeldt and Papanikolaou, 2013; Boudreau, 2010). Our objective is *prediction*, not causation.

Portfolios formed on GitHub engagement earn large and persistent spreads that remain after standard factor controls. Results are robust when we exclude Microsoft and use *harder-to-game* metrics—forks, pull requests, contributors, issues, and commits—(Appendix Tables A1 and A3). Because stars can be manipulated, we implement a fake-star cleansing procedure in the spirit

of He et al. (2024); after cleansing, the H–L returns and alphas are at least as large (Appendix Table A2). The predictive content extends to fundamentals: a present-value style decomposition shows that GitHub engagement forecasts future return on equity and intangible value added in addition to expected returns (Appendix Table A5), consistent with an organization-capital channel (Cho et al., 2024; Eisfeldt and Papanikolaou, 2013).

We further present evidence of consistency tied to platform salience. On April 14, 2020, GitHub made core features available for free. We document larger post-change return responses among firms with higher pre-change engagement (DiD-style specifications), as well as delayed post-event dynamics consistent with gradual diffusion. The delayed reaction four months after the actual policy change is evidence of initial underreaction. We interpret these patterns as *attention/visibility consistency checks*, not identification, and they are complementary to the paper's predictive focus.

Effects are most potent in software- and data-intensive industries—where open-source maps directly to production—and for engagement measures that are harder to manipulate. We therefore frame GitHub engagement as a *digital-innovation signal* most relevant in tech-intensive settings. While GitHub activity is concentrated among tech-intensive firms, these firms represent a disproportionate share of market value and innovation expenditure. Throughout, we avoid causal language and treat platform changes and ancillary exercises as robustness and economic consistency checks rather than instruments. Taken together, our findings suggest that corporate GitHub engagement contains information not fully captured by traditional financial and innovation proxies.

There is good reason to question whether investors are fully updated regarding GitHub signals. Prior studies suggest that investors tend to underreact to new information, especially when the information is complex (You and Zhang, 2009). There is room to believe that GitHub is still uncharted territory for traditional investors. First, GitHub is not part of mainstream financial data pipelines. Bloomberg, FactSet, Refinitiv, and Morningstar do not integrate GitHub metrics like stars and forks into their standard equity data feeds. This means portfolio managers would need specialized scraping procedures to monitor GitHub

activity at the firm level. Second, understanding GitHub engagement requires knowledge of software development communities, open-source culture, and repository quality indicators. Third, while alternative data research is growing, academic finance studies on GitHub are extremely rare compared to research on Twitter, Google search, or job postings. While these frictions point to a limited-attention mechanism and the possibility of gradual mispricing, they need not be the only explanation for return predictability. GitHub engagement may also proxy for exposure to systematic innovation risk arising from ecosystem competition, technological obsolescence, and network-dominance uncertainty. In that case, higher expected returns would reflect compensation for bearing innovation-related risk rather than pure mispricing. Our empirical analysis is designed to disentangle these channels, showing evidence consistent with both gradual information incorporation and a priced risk component.

Emery et al. (2025) find that GitHub repositories generate a private value of $849,000 on average upon announcement, using event-study methodology around repository release dates. Our paper takes a fundamentally different approach: rather than measuring the announcement effect of *releasing* repositories, we examine whether *ongoing community engagement* with existing repositories—measured by stars, forks, and other metrics—predicts cross-sectional variation in future returns. While Emery et al. (2025) measure the instantaneous private value of repository creation, our findings reveal that the longer-horizon return predictability—arising from ongoing community engagement rather than initial release—is where GitHub's unique information content resides. These are fundamentally complementary findings. Our finding of return predictability extending to five years points to a risk-based mechanism beyond short-term underreaction.

First, we contribute to research on intangible capital and its pricing by introducing a real-time, community-validated proxy for innovation effort and organizational know-how that *predicts* returns and fundamentals (Eisfeldt and Papanikolaou, 2013; Crouzet et al., 2022). Intangible capital can take many forms ranging from patents, trademarks, software, and brands to organization capital. Prescott and Visscher (1980) define organization capital as information generated within the firm that affects its production possibility set. From this perspective,

firms' GitHub sharing and stars can be interpreted as organization capital, since the code is generated within the firm and has the potential to influence the firm's future products and services. GitHub stars can also be thought of as components of developer ecosystem value (Boudreau, 2010). In our context, this ecosystem value feeds into organization capital since firms can focus their efforts on their projects with the highest number of stars. In yet another dimension, firms' GitHub sharing can be construed as open innovation signaling. Lerner and Tirole (2002) theorize that firms use open-source contributions to signal technical strength, innovativeness, and openness to investors, customers, and potential employees. Hence, it is difficult to categorize GitHub sharing as a single type of intangible capital; we propose that it serves as a source of innovation value that also receives limited investor attention.

Second, we add to the limited-attention and gradual-diffusion literature by showing that a nontraditional, public but cognitively costly information channel—open-source code sharing—translates into delayed price discovery and forecastable returns (Hong and Stein, 1999; Hsu et al., 2022; Hirshleifer et al., 2013). Hsu et al. (2022) demonstrate that investors underreact to new trademark registrations. Other papers have documented return predictability for firms' innovative activities measured by R&D intensity (Eberhart et al., 2004), R&D abilities (Hirshleifer et al., 2013), and patents (Fitzgerald et al., 2021). The main interpretation of these patterns is that stock prices underreact to the information contained in these activities. Because most firms do not disclose detailed data on ongoing software development and open-source contributions, our paper adds to this literature by showing that GitHub engagement contains useful information about a firm's fundamentals and is often undervalued by investors. Even in a domain where the underlying code is publicly accessible, investors may still fail to fully process the information due to the technical complexity of open-source projects, the diversity of programming languages, and uncertainty about their commercial relevance.

Third, we speak to alternative-data asset pricing by demonstrating that a widely used platform outside investor data pipes can systematically enhance cross-sectional return prediction over horizons up to five years (Green and Zhang, 2024; Dessaint et al., 2024). Alternative data in finance comprises social media (Dessaint et al., 2024), employer ratings (Green et al., 2019),

satellite images (Katona et al., 2025), and mobile devices (Froot et al., 2017) to name a few. Green and Zhang (2024) propose regulation, the growing importance of intangible assets, and vastly increased amounts of data as drivers of investors' demand in alternative data. Dessaint et al. (2024) point out that most studies employing alternative data find such data are useful to predict firms' stock returns over the short term (one year or less) while there is little evidence that it is useful to predict long-term outcomes. Our study finds, however, that firms' GitHub sharing predicts stock returns up to five years. The long-term predictability of returns points to a potential risk-based explanation (Hsu et al., 2022) in addition to the underreaction/mispricing evidence. Our paper also contributes to the literature on alternative data and the financial relevance of digital platforms (Schreieck et al., 2024) and speaks to broader questions about how open-source participation intersects with firm strategy and investor perceptions in modern capital markets (West and Gallagher, 2006; Conti et al., 2025).

Finally, we document a distinctive temporal signature in GitHub-based return predictability that reveals the production-side nature of open-source innovation. Unlike traditional innovation proxies—R&D expenditures, patents, trademarks—which predict returns at short horizons consistent with their role as disclosure events, GitHub engagement generates no predictability at lag 1 but peaks at lag 5, the only horizon at which all four return benchmarks are individually significant. This delayed emergence is consistent with a diffusion pipeline in which code development must traverse integration, testing, deployment, and adoption stages before affecting cash flows and valuations. The contrast between GitHub's delayed predictability and the immediate predictability of traditional proxies provides new evidence that open-source engagement captures a fundamentally different dimension of innovation from existing measures.

The rest of the paper proceeds as follows. Section 2 develops a simple model of open-source signaling, heterogeneous attention, and predictability, and derives testable predictions. Section 3 describes the data and measures. Section 4 presents the return and fundamentals predictability results and risk-based tests. Section 5 concludes.

2 A Theory of Open-Source Signals, Network Effects, and Predictability

We develop a framework that rationalizes why GitHub engagement should forecast returns and fundamentals. The model combines three ingredients: (i) network externalities in open-source ecosystems that amplify firm value, (ii) type-dependent sharing that generates informative signals, and (iii) heterogeneous investor attention that produces underreaction and predictability.

2.1 Open-Source Ecosystems and Network Externalities

A distinguishing feature of open-source platforms is the presence of strong network externalities. When a firm releases code on GitHub, it does not merely disclose information—it initiates a dynamic process of ecosystem development that can compound over time.

Consider a firm's repository with initial quality q_0. Community engagement generates contributions (pull requests, issues, forks) that improve the repository's functionality. Let n_t denote the number of active users at time t. The repository's effective quality evolves according to:

$$q_{t+1} = q_t + \gamma(n_t) + \eta_{t+1}, \tag{1}$$

where $\gamma(\cdot)$ is positive, increasing, and convex in n, capturing the idea that larger communities generate disproportionately more improvements through complementary modules, bug fixes, and extensions. The noise η_{t+1} represents idiosyncratic shocks to development.

User adoption follows a network externality process:

$$n_{t+1} = n_t \cdot \psi(q_t, n_t), \quad \text{where } \psi_q > 0, \psi_n > 0. \tag{2}$$

The condition $\psi_n > 0$ captures direct network effects: a repository is more valuable when others use it because of shared knowledge, compatible extensions, and labor market signaling for developers.[1]

Equations (1) and (2) generate a feedback loop: higher quality attracts users, more users improve quality, which attracts more users. This implies that *early engagement signals contain information about long-run value that compounds nonlinearly.* A repository that accumulates 1,000 stars in its first year is not merely twice as valuable as one with 500 stars—it may be on a trajectory toward ecosystem dominance.

2.2 Firm Value and Organization Capital

Let V_t denote firm value. We model the firm as owning a portfolio of repositories $\{j\}$ with associated ecosystem values $\{E_{jt}\}$. Firm value has three components:

$$V_t = A_t + \sum_j E_{jt} + \Omega_t, \tag{3}$$

where A_t represents traditional assets, E_{jt} is the ecosystem value of repository j, and Ω_t captures organization capital—the tacit knowledge, developer networks, and reputational capital that accumulate through open-source participation (Eisfeldt and Papanikolaou, 2013; Prescott and Visscher, 1980).

The ecosystem value E_{jt} depends on current engagement metrics k_{jt} (stars, forks, contributors) and expected future growth:

$$E_{jt} = f(k_{jt}) + \mathbb{E}_t \left[\sum_{s=1}^{\infty} \rho^s \Delta E_{j,t+s} \right], \tag{4}$$

where $f(\cdot)$ is increasing. Because of the feedback loop in (1)–(2), the expected growth term is convex in current engagement: repositories with high k_{jt} are more likely to achieve ecosystem dominance, generating option-like payoffs.

[1] A developer's expertise in PyTorch is more valuable when PyTorch is widely adopted, creating incentives for both firms and individuals to coordinate on dominant platforms.

Organizational capital Ω_t also responds to GitHub activity. Firms that successfully cultivate open-source communities accumulate:

- *Developer networks*: Access to talent pools who contribute to and are familiar with the firm's codebase;
- *Reputational capital*: Credibility as a technology leader that facilitates partnerships and customer acquisition;
- *Absorptive capacity*: Enhanced ability to identify and integrate external innovations (Cohen and Levinthal, 1990).

2.3 Type-Dependent Sharing and Signaling

Firms differ in their latent technological capability $\theta \in \mathbb{R}$, which maps to long-run productivity and the quality of code they can produce. Each period, the firm chooses sharing intensity $s \geq 0$. Sharing generates benefits $B(\theta, s)$—ecosystem adoption, recruiting, reputation—and costs $C(\theta, s)$—IP leakage, coordination, competitive exposure.

We assume a single-crossing condition:

$$\frac{\partial^2 B}{\partial \theta \partial s} - \frac{\partial^2 C}{\partial \theta \partial s} > 0. \tag{5}$$

High-θ firms have lower marginal costs of sharing (their code is more defensible, their competitive position more secure) and higher marginal benefits (their contributions attract more engagement, their signaling is more credible). In equilibrium, $s^*(\theta)$ is increasing in θ.

The market observes community metrics k (stars, forks, pull requests):

$$k = \mu(\theta, s^*(\theta)) + \varepsilon, \tag{6}$$

with $\mu_\theta > 0$, $\mu_s > 0$, so k is informative about latent capability. The noise ε captures platform-specific frictions and potential manipulation (e.g., fake stars). Metrics that are costlier to

fake—pull requests, contributors, commits—have lower Var(ε) and should exhibit stronger predictability.

2.4 Limited Attention and Gradual Price Discovery

If GitHub metrics were costless to process, prices would immediately reflect the information in k. However, processing GitHub signals requires:

1. *Data engineering*: Mapping firms to GitHub organizations, scraping repository data, and constructing engagement measures;

2. *Technical expertise*: Evaluating code quality, understanding programming language trends, and assessing commercial relevance;

3. *Ecosystem knowledge*: Understanding network dynamics, developer community norms, and platform-specific signals.

These processing costs create heterogeneous attention. Let $\phi \in (0, 1)$ denote the fraction of investors who process GitHub signals with precision τ. The remaining $1 - \phi$ investors ignore GitHub initially and learn gradually through indirect channels (media coverage, analyst reports, competitor announcements), consistent with Hong and Stein (1999).

Prices aggregate heterogeneous beliefs:

$$P_t = \phi\, \mathbb{E}[\theta|k, I_t] + (1 - \phi)\, \mathbb{E}[\theta|I_t], \tag{7}$$

where I_t denotes standard financial information. Define the attention wedge:

$$\Delta_t(k) \equiv \mathbb{E}[\theta|k, I_t] - \mathbb{E}[\theta|I_t]. \tag{8}$$

Then returns from t to $t + h$ satisfy:

$$P_{t+h} - P_t \propto \phi_{t+h}\Delta_{t+h}(k) - \phi_t\Delta_t(k). \tag{9}$$

Returns are predictable from k as long as the wedge closes over time—either because inattentive investors gradually learn (ϕ increases) or because the signal becomes more precise (τ increases).

The *network externality channel* amplifies this effect. Because ecosystem value compounds nonlinearly, the true $\Delta_t(k)$ may be larger than even attentive investors recognize. Stars today predict not just current quality but future ecosystem dominance—a convex, option-like payoff that is difficult to value and likely to be underweighted.

2.5 Testable Predictions

The framework generates predictions that map to our empirical tests:

Prediction 1 (Cross-Sectional Return Predictability)**.** *Higher GitHub engagement k predicts higher future returns $r_{t+1:t+H}$ because prices under-incorporate ecosystem value and network externalities. Effects should persist over multi-year horizons as the ecosystem compounds and information diffuses.*

Empirical mapping: Positive H–L spreads (Table 4) persisting to 60 months (Table 5); lagged Fama–MacBeth coefficients with delayed significance (Table 6).

Prediction 2 (Signal Quality Gradient)**.** *Predictability is stronger for metrics with lower $Var(\varepsilon)$: hard-to-fake measures (forks, pull requests, contributors) and fake-star-cleansed data.*

Empirical mapping: Appendix Tables A2 and A3.

Prediction 3 (Industry Concentration)**.** *Predictability concentrates in industries where open-source maps directly to production (software, data-intensive sectors) because μ_θ is larger and network externalities are stronger.*

Empirical mapping: Industry distribution in Table 2.

Prediction 4 (Lag Structure)**.** *Because ecosystem value compounds through feedback between quality improvements and user adoption (equations 1–2), GitHub-based return predictability should emerge at longer horizons than disclosure-based innovation signals. Patent grants, R&D*

disclosures, and trademark registrations are discrete information events that markets can process relatively quickly. In contrast, the value of open-source engagement materializes gradually as code is integrated, deployed, adopted by customers, and eventually reflected in financial reporting. We therefore predict that GitHub predictability peaks at intermediate lags (approximately 4-6 months), whereas traditional innovation proxies exhibit strongest predictability at shorter horizons.

Empirical mapping: Lag profile in Figure 1.

Prediction 5 (Fundamental Linkage)**.** *Because $s^*(\theta)$ is increasing in θ, and because ecosystem value feeds into organizational capital, higher k forecasts future ROE and intangible value added—not just expected returns.*

Empirical mapping: Decomposition results in Appendix Table A5.

Prediction 6 (Platform Salience)**.** *Events that increase platform visibility (e.g., GitHub's 2020 policy change) raise ϕ, amplifying market responsiveness to GitHub signals among previously-engaged firms.*

Empirical mapping: DiD estimates in Appendix Tables A8–A10.

Prediction 7 (Long-Horizon Persistence and Risk)**.** *The network externality channel implies that GitHub engagement proxies for exposure to ecosystem risk—the possibility that a firm's technology becomes the industry standard (positive) or is displaced by a competitor's ecosystem (negative). This systematic component implies that some portion of the return spread reflects risk compensation rather than pure mispricing.*

Empirical mapping: Factor spanning tests showing residual alpha alongside priced risk (Appendix Table A4); long-horizon persistence distinguishing risk from short-lived mispricing (Table 5).

We emphasize that our contribution is *predictive*, not causal. We demonstrate that GitHub engagement is an economically interpretable, out-of-pipeline predictor of returns and fundamentals, consistent with network externalities, organizational capital accumulation, and limited investor attention. The model provides economic intuition; it does not claim that GitHub activity *causes* value creation independent of underlying firm quality.

3 Data and Methodology

3.1 Data

We manually collect the GitHub homepage IDs of S&P 1500 companies by searching for official corporate GitHub accounts. We verify matches using company names, official websites linked from GitHub profiles, and cross-referencing with corporate press releases that reference GitHub activity. This process yields 659 unique firm-GitHub mappings covering the period 2010–2022.

Then, we scrape all repositories shared on the homepage of each company. For each repository, we collect the number of stars the repository has received. Starring a repository is a way for users to show their appreciation for a project (Borges and Valente (2018)). It is similar to liking something on social media. The number of stars a repository has received can be thought of as an indicator of its popularity and how much the community values it. In the second step, we aggregate the total number of stars for all repositories created by a single company within a month.

3.2 Methodology

For portfolio analysis, we follow Hsu et al. (2022) and form portfolios based on a size-adjusted composite measure of GitHub engagement. In each month, we scale firms' GitHub activity metrics—stars, forks, watchers, and repository counts—by lagged market equity and extract the first principal component as the summary measure of GitHub engagement. Firms are then sorted into low (L), middle (M), and high (H) portfolios according to the 33rd and 67th percentiles of this measure. We further construct a high-minus-low portfolio that takes a long (short) position in the high (low) GitHub engagement portfolio.

For the decomposition analysis, Cho et al. (2024) break down the market-to-book ratio into the components of stock return, return on equity, and investment value added. Their conjecture is that a return predictor needs to be linked to future profitability, investment, or discount rates. We adopt this approach to explain the number of stars a firm's GitHub repository has received. Specifically, we decompose the logarithm of a firm's GitHub stars into contributions

from future discount rates, returns on equity, and the value added from future value-enhancing investment in or divestment from book equity.

$$stars_t \approx -\sum_{j=1}^{\infty} \rho^{j-1} r_{t+j} + \sum_{j=1}^{\infty} \rho^{j-1} roe_{t+j} + \sum_{j=1}^{\infty} \rho^{j-1} iva_{t+j} \tag{10}$$

where each variable on the left-hand side and right-hand side represents the natural logarithm of the corresponding original variable. We compute the discounted sums using realized variables over 2-, 4-, and 6-year as well as infinite horizons. We form three portfolios, sorted by GitHub stars, and compute the discounted sum of the relevant variables at the portfolio level using a discount factor $\rho = 0.96$ (Cho et al. (2024)). We then regress these sums on the portfolio's current GitHub stars with year fixed effects. We estimate

$$\sum_{j}^{J} \rho^{j-1} y_{i,t+j} = \alpha_t + \beta stars_{i,t} + \varepsilon_{i,t} \tag{11}$$

for $y \in \{roe, iva, r\}$ so that β estimates each of the three components of the decomposition.

4 Empirical Results

This section presents the empirical findings on the relationship between firms' GitHub activity and various firm characteristics and stock market outcomes. We begin by describing the distribution of GitHub activity across firms and industries, then examine cross-sectional differences in firm characteristics based on GitHub stars, and finally assess the return predictability associated with GitHub activity through portfolio analysis and regression frameworks.

4.1 Descriptive Statistics

Table 1 reports the top ten companies with the highest GitHub engagement. Large technology firms such as Microsoft, IBM, Alphabet, and Intel dominate the list, indicating a concentration of GitHub activity among established tech-oriented firms.

Table 2 shows the distribution of GitHub engagement across Fama-French 10 industries from 2009 to 2022. The Business Equipment category, including computers, software, and electronic equipment, exhibits the highest level of GitHub activity, particularly after 2012, with a dramatic rise observed from 2014 onward. The Other category, which includes mining, construction, transportation, hotels, business services, entertainment, and finance, comes a distant second. Certain other industries, such as Manufacturing, Consumer Nondurables, and Healthcare, also show increasing participation, albeit at lower levels. This pattern suggests the diffusion of open-source contributions beyond core technology sectors over time.

4.2 Firm Characteristics by GitHub Star Groups

To examine the characteristics of firms with different levels of GitHub recognition, Table 3 sorts firms into terciles (Low, Middle, High) based on the number of GitHub stars received in the prior year. Firms in the high-star group tend to be larger, more innovation-intensive, and exhibit slightly lower book-to-market ratios. While the R&D ratio shows a modest increase across the terciles, the innovation measure based on patents does not display strong variation. This indicates that the GitHub stars measure is not fully subsumed by traditional innovation measures. Return skewness is also slightly higher in the high-star group. These results suggest that GitHub popularity is associated with firm size and some innovation proxies, although not uniformly across all measures.

4.3 Portfolio Analysis

Table 4 presents the baseline portfolio evidence using a size-adjusted composite measure of GitHub engagement. A potential concern with raw GitHub activity measures is that they may mechanically correlate with firm size since larger firms tend to operate more repositories and attract greater attention from the developer community. To address this concern, we construct a size-adjusted GitHub engagement measure using a principal component analysis (PCA) approach. Specifically, we first scale GitHub activity metrics—stars, forks, watchers, and

the number of repositories—by the firm's lagged market equity. Forking refers to the creation of a new software repository by copying an existing repository. Prior research shows that forking is often used to contribute improvements to the original project and tends to occur more frequently for repositories in popular technological areas (Jiang et al., 2016). Users can also watch repositories to receive notifications about project activity. Watchers represent passive project participants who follow development and may later contribute to the project (Sheoran et al., 2014). We then extract the first principal component from these scaled measures using a rolling-window PCA estimated only with information available up to each month.

Each month, firms are sorted into terciles based on the resulting PCA engagement score, and value-weighted portfolios are formed and held for one month. The high-minus-low (H-L) portfolio generates an economically and statistically significant excess return of 1.45% per month (t = 2.91) and an industry-adjusted return of 0.88% (t = 2.44). These return differentials are not fully explained by the Fama-French five factor model or the q-factor model, with alphas of 1.55% and 1.72%, respectively. All reported t-statistics in the portfolio analyses are computed using Newey and West (1987) standard errors with six lags, which account for heteroskedasticity and time-series dependence in monthly portfolio returns. The persistence of the return spread, combined with modest R^2 values, suggests that GitHub activity contains incremental information not captured by standard risk factors. These results indicate that the return predictability associated with GitHub engagement is not driven by firm size. Instead, the size-scaled PCA measure captures cross-sectional variation in firms' open-source activity relative to their economic scale, suggesting that firms whose GitHub engagement is disproportionately high relative to their market capitalization earn higher subsequent returns.

As shown in Table 1, Microsoft dominates the list of firms with the highest number of GitHub repository shares. In addition, Microsoft acquired GitHub for $7.5 billion in October 2018. To remove concerns that our results are driven by Microsoft, we conduct a robustness check where we exclude Microsoft from our sample. Results are displayed in Appendix Table A1. There are no meaningful differences when we include or exclude Microsoft. Consistent with the baseline results, the H-L portfolio yields economically and statistically significant excess

returns that are not fully explained by the Fama-French five-factor or the q-factor models. These results indicate that GitHub sharing is associated with positive abnormal returns not just for the company that owns the platform but for any other company that shares on the platform.

He et al. (2024) highlight the risk of GitHub stars being artificially inflated by bots, which has become a more prominent issue since 2024. We implement the authors' *StarScout* algorithm to detect fake stars. Appendix Table A2 removes suspected fake stars and runs the star-based portfolio tests as a robustness check. Results are similar, and the t-statistics for the spreads and alphas are slightly larger. He et al. (2024) also point out that unlike stars, there are other popularity and engagement metrics in GitHub that are hard to fake. Some examples are the number of subscribers, contributors, pull requests, issues, and commits.[2] Appendix Table A3 provides a summary table for the results using these hard-to-fake engagement metrics. High-minus-Low return spreads and alphas are all significant, allaying concerns that the return predictability of stars is artificially generated.

Table 5 extends the analysis by examining the performance of the H-L portfolio over longer horizons. The excess return compounds to 15.25% over 12 months and reaches 52.48% over 60 months, with strong and persistent t-statistics. Alphas from both the five-factor and q-factor models remain positive and significant, suggesting that GitHub activity captures a persistent source of return heterogeneity. The increasing t-statistics over time reflect a compounding pattern consistent with the network externalities highlighted in the theoretical framework, in contrast to the decaying predictability typical of anomaly strategies. The fact that excess returns persist over a long period of time is also suggestive of a systematic risk-based explanation (Hsu et al. (2022)).

[2]A pull request is a way for someone to propose changes to a codebase so they can be reviewed, discussed, and possibly merged into the main project. An issue is a record used to track and discuss a specific task, bug, feature request, or question in a repository. Issues focus on what needs to be done or addressed whereas pull requests focus on how the proposed solution is implemented and integrated into the codebase. Finally, a commit is a recorded snapshot of changes made to files in a repository at a specific point in time.

4.4 Lag Structure

Table 4 illustrates the return predictability of GitHub engagement measures over the next month. Figure 1 examines return predictability from the concurrent month until ten months into the future. Panel A shows the lag profile covering all 11 lags from 0 through 10. Each line shows the t-stat associated with one of four coefficients: Excess Return, Industry-Adjusted Return, FF5 Alpha, and HXZ Alpha. The pattern is striking. Lag 0 yields three of four coefficients significant, indicating that a contemporaneous signal exists. Lag 1 drops to zero significant coefficients, lags 2–3 produce two to three significant coefficients, lag 4 dips to zero, lag 5 achieves significance in all four benchmarks (the unique peak), and lags 6–10 decline.

The lag 0 to lag 1 to lag 5 progression implies that the market observes GitHub activity contemporaneously but the signal disappears at lag 1, consistent with investors initially being unable to incorporate it into prices, then resurfaces at lag 5 when the production-side value materializes.

Appendix Table A6 and Table 8 present additional evidence on the lag structure of GitHub stars and how it compares with the lag structure of traditional innovation proxies. For example, GitHub PCA goes from -0.55% (t = -0.72) at lag 1 to 0.77% (t = 2.08) at lag 5, while patents go from 1.40% (t = 1.92) at lag 1 to 0.31% (t = 0.89) at lag 5.

Figure 1 Panel C shows the monotonicity of the results, that is, whether the High engagement group earns higher returns than the Middle group and the Middle group earns higher returns than the Low group. It shows that this return ordering is monotonically preserved at 9 out of 11 lags.

4.5 Cross-Sectional and Panel Regressions

Table 6 reports Fama-MacBeth cross-sectional regressions examining whether GitHub engagement predicts subsequent stock returns after controlling for firm characteristics. Across specifications, the coefficients on $ln(1 + watchers)$, $ln(1 + stars)$, $ln(1 + forks)$, and PC1, the first principal component, are all positive and statistically significant at the 1% level. The

economic interpretation is straightforward: firms with higher GitHub engagement earn higher future returns, even after controlling for firm size, book-to-market ratio, R&D intensity, and traditional innovation measures.

Importantly, PC1—the first principal component extracted from the rolling-window PCA of GitHub engagement metrics—also exhibits a positive and highly significant coefficient. This result indicates that the predictive content is not driven by any single popularity metric but instead reflects a common engagement factor embedded in open-source activity.

The robustness of the GitHub coefficients in the presence of controls and fixed effects confirms that the documented return predictability is incremental to firm size and conventional innovation proxies. These findings are consistent with the limited-attention mechanism outlined in Section 2.4, whereby technically complex and out-of-pipeline signals are incorporated into prices gradually.

Table 7 reports panel fixed-effects regressions that examine within-firm variation in returns associated with changes in GitHub engagement. The inclusion of firm fixed effects controls for time-invariant firm characteristics, while year-month fixed effects absorb aggregate shocks. Across all specifications, the coefficients on $ln(1 + watchers)$, $ln(1 + stars)$, $ln(1 + forks)$, and PC1 are all positive and statistically significant at the 1% level. The magnitudes are economically substantive and suggest that increases in GitHub engagement within a firm are associated with higher subsequent returns.

Taken together, the Fama-MacBeth and panel regressions provide complementary evidence. The former demonstrates cross-sectional predictability, while the latter shows that the relation also holds within firms over time. This dual evidence strengthens the conclusion that GitHub engagement contains economically relevant information beyond traditional innovation proxies.

4.6 Factor Pricing Analysis

The portfolio and regression analyses establish that GitHub engagement predicts future returns. A natural question is whether this predictability reflects compensation for systematic risk exposure or represents mispricing that should be arbitraged away. If GitHub engagement

proxies for exposure to priced risk factors—for instance, uncertainty about technological ecosystem outcomes—then the return spread would represent fair compensation rather than an anomaly.

To investigate this possibility, we follow the cross-sectional regression framework of Hou et al. (2020). We construct a GitHub-based factor as the return on a portfolio that is long firms in the top tercile of GitHub stars and short firms in the bottom tercile. We then test whether loadings on this factor help explain cross-sectional variation in expected returns across a broad set of test assets, controlling for established risk factors.

Appendix Table A4 reports the results. Panel A uses single-sorted portfolios as test assets; Panel B uses long-minus-short anomaly portfolios. We regress average portfolio returns on factor betas, including the Fama and French (2018) six factors, the augmented q-factor model of Hou et al. (2021), and behavioral factors from Daniel et al. (2020).

The GitHub factor exhibits significant positive pricing power across specifications. In Panel A, the GitHub factor commands a monthly risk premium of 1.76–2.05% when included alongside standard factors. This premium remains significant even when controlling for both traditional and behavioral factors. Panel B shows similar patterns, with premia ranging from 3.58% to 5.21% per month.

These results suggest that cross-sectional variation in GitHub engagement is associated with priced risk. One interpretation is that firms with high GitHub activity are exposed to "ecosystem risk"—the systematic uncertainty about whether a firm's open-source investments will achieve network dominance or be displaced by competing ecosystems. Such risk would rationally command a premium. The factor pricing results thus provide evidence that GitHub engagement captures exposure to innovation-related systematic risk beyond what existing factor models span.

4.7 Decomposition of GitHub Stars

The preceding analyses document return predictability but do not reveal its economic sources. Cho et al. (2024) show that any variable related to firm valuation can be decomposed into

contributions from three channels: expected future profitability (return on equity), expected returns (discount rates), and investment value added. Their key insight is that *any cross-sectional return predictor must have a mirror image in cash-flow fundamentals*—the predictor cannot forecast returns without also forecasting profitability, investment, or changes in valuation ratios.

We apply this framework to understand the economic content of GitHub stars. Specifically, does GitHub engagement predict returns because high-star firms have (i) superior future operating performance, (ii) higher discount rates reflecting systematic risk, or (iii) more valuable future investment opportunities?

Following Cho et al. (2024), we decompose the log of GitHub stars using the present-value identity:

$$\text{stars}_{i,t} \approx \sum_{j=1}^{J} \rho^{j-1} \text{roe}_{i,t+j} - \sum_{j=1}^{J} \rho^{j-1} r_{i,t+j} + \sum_{j=1}^{J} \rho^{j-1} \text{iva}_{i,t+j} \tag{12}$$

where roe is return on equity (profitability), r is stock returns (discount rates), iva is investment value added, and $\rho = 0.96$. To implement this, we sort firms into terciles by GitHub stars, compute discounted sums of realized values for each component at the portfolio level over 2-, 4-, and 6-year horizons, and regress each component on current GitHub stars with year fixed effects.

Appendix Table A5 reports the decomposition results. Over a six-year horizon, the ROE coefficient is 0.324. Firms with higher GitHub stars subsequently earn higher returns on equity. This is the dominant channel—the profitability coefficient is approximately twice as large as the expected return coefficient. The expected return coefficient is 0.163. Higher-star firms also have higher realized returns, consistent with compensation for risk exposure documented in Section 4.6. Finally, the IVA coefficient is 0.107. GitHub engagement predicts future value-enhancing investment, suggesting that high-star firms continue making investments that the market values positively.

The decomposition reveals that GitHub stars contain genuine information about firm fundamentals, not merely investor sentiment. Three findings stand out. First, the profitability channel

dominates. The ROE coefficient (0.324) is roughly twice the expected return coefficient (0.163), indicating that most of the GitHub-return relation reflects information about future cash flows rather than discount rates. Firms with active open-source engagement subsequently generate superior operating performance—consistent with GitHub activity proxying for organization capital that takes time to manifest in accounting earnings. This pattern mirrors findings in the innovation literature: Hirshleifer et al. (2013) document that innovative efficiency predicts returns primarily through a cash-flow channel, and Hsu et al. (2022) find similar patterns for trademarks. Second, the non-trivial expected return coefficient supports the risk interpretation from Section 4.6. Approximately one-third of the total signal (0.163 out of $0.324 + 0.163 \approx 33\%$) reflects compensation for bearing innovation risk, not mispricing. This helps explain why return predictability persists over long horizons (Table 5)—pure mispricing would typically be arbitraged away more quickly, but risk compensation can persist indefinitely.

Third, the positive IVA coefficient indicates that GitHub engagement predicts future value creation through investment. Firms with successful open-source ecosystems continue to identify and execute valuable projects, suggesting persistent organizational capabilities rather than one-time advantages.

We attempt to reconcile risk and mispricing. The decomposition results are consistent with *both* limited attention and risk-based explanations operating simultaneously. The dominance of the profitability channel suggests that investors underweight publicly available GitHub signals when forecasting future earnings—consistent with the gradual-diffusion mechanism in our theoretical framework. At the same time, the substantial expected return coefficient indicates that some return spread reflects fair compensation for ecosystem risk. These findings parallel Hsu et al. (2022), who conclude that trademark return predictability reflects both underreaction to future fundamentals and exposure to systematic risk.

4.8 Comparison with Traditional Innovation Proxies

This section benchmarks GitHub-based return predictability against conventional innovation measures used in the asset-pricing and innovation literatures. Table 8 is complementary to

earlier evidence in two ways. First, Table 4 documents one-month-ahead return predictability from sorting on individual GitHub measures. Second, Table 5 studies the same high–minus–low GitHub strategy over *longer holding horizons* (12–60 months). In contrast, Table 8 focuses on a distinct question: *when we use the same portfolio laboratory and the same inference, how does GitHub compare to standard innovation proxies such as R&D/ME, patents, and trademarks?* This apples-to-apples comparison is important for interpretation and for positioning our contribution relative to the existing innovation-proxy literature.

Table 8 reports equal- and value-weighted *monthly* decile portfolios formed on four GitHub measures—stars, forks, watchers, and repositories—as well as a GitHub composite constructed as the first principal component (PC1) from a rolling-window PCA applied to standardized versions of these four measures (PCA on the correlation matrix). Portfolios are formed at month t using signals measured at month $t-5$ and are held for one month; Table 8 therefore evaluates *one-month-ahead* return predictability. For benchmarking, the same portfolio construction is applied to R&D/ME, patent counts, and trademark counts. The table reports mean monthly excess returns (in percent), Newey–West t-statistics with six lags, and annualized Sharpe ratios for the low and high deciles and the corresponding high–low spread.

The main new insight is comparative. GitHub-based sorts generate economically meaningful long–short spreads that are of similar order of magnitude to those obtained from established innovation measures under identical portfolio construction and inference. In value-weighted portfolios, the long–short spreads for GitHub measures and for the GitHub PCA composite are comparable to the spreads produced by patents, trademarks, and R&D/ME, with moderate Sharpe ratios. At the same time, equal-weighted spreads are attenuated for GitHub signals, reflecting the discreteness and sparsity of GitHub activity (many firm-months have zero counts), which induces bunching in the bottom decile. This pattern clarifies *where* GitHub's predictive content resides: it is most apparent in the value-weighted universe where GitHub engagement is more prevalent and measured with less discretization.

A related concern is that raw GitHub counts may partly proxy for firm scale. Table 4 addresses this by repeating the analysis after scaling the count-based measures by lagged

market equity (ME). Specifically, we form scaled GitHub components (stars/ME, forks/ME, watchers/ME, repositories/ME) and construct a scaled GitHub composite as PC1 from a rolling-window PCA on standardized scaled components. For comparability, we also scale patent and trademark counts by ME, while R&D/ME is already a ratio. The scaled results confirm that GitHub-based long–short spreads remain economically meaningful and broadly comparable to the traditional innovation benchmarks, indicating that the Table 8 findings are not driven purely by a mechanical size channel.

The five-month lag is motivated by an ex-ante timing prior rather than ex-post tuning. GitHub engagement is a production-side signal that must traverse integration and QA, release management, customer adoption, and reporting/attention cycles before affecting cash flows and valuation. This diffusion pipeline implies a natural delay on the order of one to two quarters, making a mid-lag window (approximately 4–6 months) economically plausible. We therefore use a five-month lag as a representative point within this window. To guard against concerns about cherry-picking, the Appendix reports lag-profile evidence over a pre-specified family of lags and shows that GitHub predictability is concentrated at mid-lag horizons, whereas traditional innovation proxies behave more like disclosure/news variables with predictive content concentrated at shorter horizons. Appendix Table A6 introduces a new composite GitHub activity measure, open-source activity index (OSAI). OSAI aggregates multiple hard-to-fake activity measures—forks, pull requests, contributors, issues, and commits—that directly reflect sustained developer effort and community interaction. For each firm-month, we log-transform these activity counts, residualize them with respect to firm size, standardize the residuals cross-sectionally, and take their simple average. This construction mitigates mechanical size effects and reduces noise inherent in any single GitHub metric, yielding a cleaner proxy for organization capital embedded in open-source production.

We further construct a peer-adjusted OSAI measure to isolate GitHub engagement that is incremental relative to firms with similar observable innovation characteristics. Specifically, we demean OSAI within size-by-R&D-intensity peer groups formed each month, removing broad-level differences associated with firm scale and conventional innovation investment. This

adjustment ensures that the signal reflects within-peer variation in open-source engagement rather than systematic differences across firm types. Together, OSAI and its peer-adjusted counterpart allow us to assess whether GitHub-based measures contain return-relevant information beyond both raw popularity indicators and standard innovation proxies such as R&D intensity, patents, and trademarks.

4.9 Out-of-Sample Analysis

While the portfolio evidence in Sections 4.3–4.6 and Section 4.8 evaluates return predictability through cross-sectional sorting strategies, Table 9 asks a stricter and complementary question: *does a GitHub-based signal improve statistical return forecasts when the model is re-estimated in real time and evaluated strictly out of sample?* This exercise matters for two reasons. First, it addresses the concern that in-sample predictability may reflect sampling variation, overfitting, or implicit look-ahead. Second, it places GitHub-based measures and traditional innovation proxies on a common forecasting metric that is widely used in the return-prediction literature, namely the out-of-sample R^2 relative to a simple historical-mean benchmark (**?**).

Forecasting setup. For each signal $x_{i,t}$, we form a pooled predictive regression for monthly stock returns using only information available at the time of forecast formation. Let $r_{i,t}$ denote the excess return of stock i in month t. For a given lag $k \in \{1, 5\}$, we estimate the following pooled OLS regression using an expanding window that ends at $t-1$:

$$r_{i,\tau} = \alpha_{t-1} + \beta_{t-1}\, x_{i,\tau-k} + \varepsilon_{i,\tau}, \qquad \tau \le t-1. \tag{13}$$

We then generate a one-step-ahead forecast for month t:

$$\widehat{r}_{i,t|t-1} = \widehat{\alpha}_{t-1} + \widehat{\beta}_{t-1}\, x_{i,t-k}. \tag{14}$$

As a benchmark, we use the pooled historical-mean (intercept-only) forecast estimated on the same expanding window:

$$\widehat{r}^{bm}_{i,t|t-1} = \widehat{\mu}_{t-1}, \tag{15}$$

where $\widehat{\mu}_{t-1}$ is the mean of $r_{i,\tau}$ over the estimation sample $\tau \leq t-1$. We evaluate forecast accuracy over the out-of-sample period using mean squared forecast errors:

$$\text{MSE}_{model} = \frac{1}{N_{\text{OOS}}} \sum_{(i,t)\in\text{OOS}} \left(r_{i,t} - \widehat{r}_{i,t|t-1}\right)^2, \qquad \text{MSE}_{bm} = \frac{1}{N_{\text{OOS}}} \sum_{(i,t)\in\text{OOS}} \left(r_{i,t} - \widehat{r}^{bm}_{i,t|t-1}\right)^2. \tag{16}$$

The out-of-sample fit statistic reported in Table 9 is:

$$R^2_{\text{OOS}} = 1 - \frac{\text{MSE}_{model}}{\text{MSE}_{bm}}. \tag{17}$$

We compute R^2_{OOS} for the full sample ("Overall") and separately within size terciles, and we report results for lag-1 and lag-5 specifications.

A central feature of Table 9 is that all R^2_{OOS} values are extremely small, typically on the order of 10^{-4} to 10^{-3}. This is not surprising: forecasting *individual-stock* monthly returns is notoriously difficult, and even economically meaningful predictors often deliver only tiny reductions in forecast MSE relative to a constant benchmark. Accordingly, small positive values indicate modest improvements in forecasting accuracy, while small negative values indicate slight underperformance relative to the pooled-mean benchmark.

The first takeaway is that GitHub-based signals do not deliver large unconditional gains in out-of-sample forecasting accuracy—but neither do traditional innovation proxies. For example, at the one-month horizon (lag 1), the GitHub measures produce slightly negative R^2_{OOS} overall and across most size buckets, while R&D/ME and patents exhibit small positives in some terciles and trademarks are mostly negative. At the five-month horizon (lag 5), the GitHub signal shows pockets of improvement: OSAI (lag 5) is slightly positive in the overall sample ($R^2_{\text{OOS}} \approx 0.00013$) and notably positive for the smallest size tercile, while remaining slightly negative for the middle and largest terciles. The peer-adjusted signal exhibits a similar pattern

with improvement concentrated among smaller firms. The conventional innovation measures display comparably mixed results across size terciles.

The main new insight from this section is therefore *comparative and methodological*. Methodologically, the expanding-window design ensures that the forecasting exercise is strictly real-time and guards against look-ahead. Comparatively, Table 9 shows that GitHub-based measures perform *at least as well as* standard innovation proxies in a stringent out-of-sample evaluation, even though all predictors yield only modest MSE improvements at the monthly horizon. Taken together with the portfolio evidence, these findings suggest that GitHub engagement contains economically relevant information for expected returns, but that the incremental forecasting gains in unconditional one-step-ahead return prediction are necessarily small in this difficult setting.

As an additional robustness check, we examine whether GitHub engagement improves out-of-sample return prediction in a flexible nonlinear forecasting framework. Specifically, we estimate a long short-term memory (LSTM) model that uses past returns and standard Fama-French factors as baseline inputs, and then augment the model with GitHub engagement measured by the firm-level star count. The model is trained and evaluated strictly out of sample using a rolling forecasting design. Table A7 reports the results. Relative to the baseline specification, adding GitHub engagement modestly reduces out-of-sample mean squared error and increases out-of-sample R^2. While the incremental gains are economically small—as is typical in monthly return forecasting—these results indicate that GitHub engagement contains predictive information that is not fully captured by standard factor inputs, even in a nonlinear learning environment. We note that the R^2 reported in Table A7 reflects the model's overall explanatory power in a cross-sectional forecasting framework and is therefore not directly comparable to the R^2_{OOS} values in Table 9, which measure incremental improvement over a historical-mean benchmark and are naturally much smaller in magnitude.

4.10 Difference-in-Differences Analysis

Appendix Table A8 presents a difference-in-differences (DiD) analysis exploiting GitHub's policy change on April 14, 2020, which made core features freely available to all users. This change effectively lowered the barrier to participation and sharing on the platform. We define a treatment group in four different ways: In Column (1), it is firms whose GitHub watcher count was above the median in the 12 months prior to the policy shift. Subsequently, in Columns (2) to (4) we replace watchers with stars, forks, and repository count. The variable of interest is the interaction term between the post-policy period and the treatment group indicator. The timeframe for this analysis is from April 2017 to April 2023. Using all four measures, the analysis shows that treated firms experience a statistically significant increase in returns following the policy change, suggesting that greater visibility or engagement on GitHub translates into favorable investor responses. The effect is also robust to controlling for one-month lagged returns. These findings further support the interpretation that GitHub activity reflects meaningful firm-specific information that investors respond to when platform frictions are reduced.

We do not claim causality based on this analysis, nor does this analysis form our main identification test; rather, we present this analysis as evidence consistent with our mechanism.

To assess the validity of the DiD strategy underlying the estimates in Appendix Table A8, we test the parallel trends assumption by examining the dynamic pattern of returns around the treatment date. Appendix Table A9 reports the results of an event-time specification that replaces the treatment indicator in the baseline DiD regression with a set of leads and lags relative to the treatment month. Coefficients on the pre-treatment leads capture differences in returns between treated and control firms before the policy change, while post-treatment lags measure the evolution of the treatment effect over time. The general absence of statistically significant coefficients in the pre-treatment period suggests that treated and control firms followed similar return trends prior to treatment, supporting the parallel trends assumption. Statistically significant coefficients in the post-treatment period indicate that the treatment effect emerges after the policy change and persists for several months thereafter.

The presence of statistically significant coefficients, especially in month +4, points to investors' delayed reaction, potentially due to investors' lack of familiarity with GitHub. This is consistent with GitHub producing costly-to-process public signals. Even though the 2020 policy change on GitHub was publicly available information, traditional investors were unable to gauge the effects immediately since GitHub metrics are unavailable in traditional finance data feeds and most traditional investors are unfamiliar with them.

Appendix Table A10 investigates whether the treatment effect of GitHub engagement varies with the intensity of firms' pre-treatment engagement on the platform. We estimate DiD models that interact the treatment indicator with firms' GitHub engagement measures per employee. The results reveal that the post-treatment return premium is concentrated among firms with high engagement. This is especially true for stars, forks, and repository counts. These findings highlight heterogeneous effects consistent with the broader mechanisms proposed in the paper, including investor attention and enhanced perceptions of innovation capacity, whereby treatment effects are stronger for firms with greater initial visibility on the platform. All DiD results are robust to controlling for firm size.

We acknowledge that our DiD analysis coincides with the COVID-19 pandemic, which broadly affected technology stocks. However, our treatment-control comparison within the tech-intensive GitHub sample helps mitigate this concern, as both treated and control firms were exposed to similar pandemic-related shocks. The parallel trends test in Appendix Table A9 shows no significant pre-treatment divergence, supporting the validity of our design.

5 Conclusion

This paper investigates whether publicly observable GitHub activity offers valuable insights into firm value, stock return predictability, and economic fundamentals. Leveraging a novel dataset that links GitHub repositories to publicly traded U.S. firms, we show that GitHub activity metrics—such as stars, forks, and watchers—serve as informative and timely signals of corporate innovation and technological engagement.

Our empirical analyses yield several key findings. First, GitHub activity is concentrated in technology firms but is becoming increasingly prevalent across other industries. Second, firms with higher GitHub recognition tend to be larger and more innovation-intensive. Third, portfolios formed on the basis of GitHub engagement—measured using a market equity-scaled principal component of stars, forks, watchers, and repositories—generate significant abnormal returns, even after accounting for common risk factors. We ensure the robustness of our results by running portfolio tests after cleansing fake stars and using hard-to-fake engagement metrics such as forks and pull requests. Fourth, a decomposition of GitHub activity reveals that stars reflect not only expected returns but also future profitability and intangible value, suggesting that GitHub engagement captures both risk-based and fundamental components of firm valuation. Fifth, factor pricing, benchmarking against traditional innovation proxies, and out-of-sample forecasting tests indicate that GitHub engagement captures both systematic innovation risk and economically meaningful information that is not fully subsumed by existing measures. Finally, to provide consistent evidence, a difference-in-differences analysis around GitHub's 2020 policy change reveals that firms with greater pre-policy GitHub engagement experienced significantly higher returns post-change, suggesting that investors respond favorably when platform frictions are reduced and open-source signals become more visible. Investors' delayed response to GitHub policy changes is evidence of initial underreaction, as GitHub is not part of standard finance data feeds.

Together, these results underscore the financial relevance of digital platform participation and highlight the potential of alternative data, particularly open-source indicators, as complements to traditional financial and innovation metrics. More broadly, the paper contributes to the growing literature on intangible capital, alternative data in asset pricing, and the strategic use of open-source platforms in shaping market perceptions.

These findings have practical implications for multiple stakeholders. For corporate managers, active engagement with open-source communities can serve not only as a strategic innovation practice but also as a financial communication channel that signals technological capability. Publishing high-quality, publicly visible repositories may improve visibility among developers

and investors alike, potentially influencing both technological reputation and capital market outcomes. For institutional investors and quantitative researchers, GitHub metrics represent an alternative data source that remains outside standard financial data pipelines (Bloomberg, FactSet, Refinitiv, and Morningstar do not integrate GitHub metrics), creating potential alpha opportunities for those willing to invest in data engineering and technical interpretation. For academic researchers, our decomposition analysis demonstrates that GitHub stars reflect fundamental firm characteristics including future profitability and intangible value, providing a new lens for studying innovation, organization capital, and the pricing of intangible assets.

References

Borges, Hudson, and Marco T Valente, 2018, What's in a github star? understanding repository starring practices in a social coding platform, *Journal of Systems and Software* 146, 112–129.

Boudreau, Kevin, 2010, Open platform strategies and innovation: Granting access vs. devolving control, *Management Science* 56, 1849–1872.

Cho, Thummim, Lukas Kremens, Dongryeol Lee, and Christopher Polk, 2024, Scale or yield? a present-value identity, *Review of Financial Studies* 37, 950–988.

Cohen, Wesley M, and Daniel A Levinthal, 1990, Absorptive capacity: A new perspective on learning and innovation, *Administrative Science Quarterly* 35, 128–152.

Conti, Annamaria, Christian Peukert, and Maria Roche, 2025, Beefing it up for your investor? engagement with open source communities, innovation, and startup funding: Evidence from github, *Organization Science* .

Crouzet, Nicolas, Janice C Eberly, Andrea L Eisfeldt, and Dimitris Papanikolaou, 2022, The economics of intangible capital, *Journal of Economic Perspectives* 36, 29–52.

Daniel, Kent, David Hirshleifer, and Lin Sun, 2020, Short- and long-horizon behavioral factors, *Review of Financial Studies* 33, 1673–1736.

Dessaint, Olivier, Thierry Foucault, and Laurent Fresard, 2024, Does alternative data improve financial forecasting? the horizon effect, *The Journal of Finance* 79, 2237–2287.

Eberhart, Allan C, William F Maxwell, and Akhtar R Siddique, 2004, An examination of long-term abnormal stock returns and operating performance following rd increases, *The Journal of Finance* 59, 623–650.

Eisfeldt, Andrea L, and Dimitris Papanikolaou, 2013, Organization capital and the cross-section of expected returns, *The Journal of Finance* 68, 1365–1406.

Emery, Logan P, Chan Lim, and Shiwei Ye, 2025, The private value of open-source innovation SSRN Working Paper No. 5054880.

Fama, Eugene F, and Kenneth R French, 2015, A five-factor asset pricing model, *Journal of Financial Economics* 116, 1–22.

Fama, Eugene F, and Kenneth R French, 2018, Choosing factors, *Journal of Financial Economics* 128, 234–252.

Fama, Eugene F, and James D MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* 81, 607–636.

Fitzgerald, Tristan, Benjamin Balsmeier, Lee Fleming, and Gustavo Manso, 2021, Innovation search strategy and predictable returns, *Management Science* 67, 7635–7655.

Froot, Kenneth, Namho Kang, Gideon Ozik, and Ronnie Sadka, 2017, What do measures of real-time corporate sales say about earnings surprises and post-announcement returns?, *Journal of Financial Economics* 125, 143–162.

Green, T Clifton, Ruoyan Huang, Quan Wen, and Dexin Zhou, 2019, Crowdsourced employer reviews and stock returns, *Journal of Financial Economics* 134, 236–251.

Green, T Clifton, and Shaojun Zhang, 2024, Alternative data in active asset management, *Fisher College of Business Working Paper Series* .

Grossman, Sanford J, and Joseph E Stiglitz, 1980, On the impossibility of informationally efficient markets, *The American Economic Review* 70, 393–408.

Gupta, Abhinav, Naman Nishesh, and Elena Simintzi, 2024, Big data and bigger firms: A labor market channel SSRN Working Paper No. 4994837.

He, Hao, Haoqin Yang, Philipp Burckhardt, Alexandros Kapravelos, Bogdan Vasilescu, and Christian Kastner, 2024, 4.5 million (suspected) fake stars in github: A growing spiral of popularity contests, scams, and malware, *arXiv:2412.13459 [cs.CR]* .

Hirshleifer, David, Po-Hsuan Hsu, and Dongmei Li, 2013, Innovative efficiency and stock returns, *Journal of Financial Economics* 107, 632–654.

Hong, Harrison, and Jeremy C Stein, 1999, A unified theory of underreaction, momentum trading, and overreaction in asset markets, *The Journal of Finance* 54, 2143–2184.

Hou, Kewei, Haitao Mo, Chen Xue, and Lu Zhang, 2021, An augmented q-factor model with expected growth, *Review of Finance* 25, 1–41.

Hou, Kewei, Chen Xue, and Lu Zhang, 2015, Digesting anomalies: An investment approach, *Review of Financial Studies* 28, 650–705.

Hou, Kewei, Chen Xue, and Lu Zhang, 2020, Replicating anomalies, *Review of Financial Studies* 33, 2019–2133.

Hsu, Po-Hsuan, Dongmei Li, Qin Li, Siew H Teoh, and Kevin Tseng, 2022, Valuation of new trademarks, *Management Science* 68, 257–279.

Jiang, Jing, David Lo, Jiahuan He, Xin Xia, Pavneet S Kochhar, and Li Zhang, 2016, Why and how developers fork what from whom in github, *Empirical Software Engineering* 22, 547–578.

Katona, Zsolt, Marcus O Painter, Panos N Patatoukas, and Jean Zeng, 2025, On the capital market consequences of big data: Evidence from outer space, *Journal of Financial and Quantitative Analysis* 60, 551–579.

Koch, Simon, David Klein, and Martin Johns, 2024, The fault in our stars: An analysis of github stars as an importance metric for web source code, in *Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb)*.

Lerner, Josh, and Jean Tirole, 2002, Some simple economics of open source, *The Journal of Industrial Economics* 197–234.

Lerner, Josh, and Jean Tirole, 2005, The economics of technology sharing: Open source and beyond, *Journal of Economic Perspectives* 19, 99–120.

Maxfield Brown, Eva, Cailean Osborne, Peter Cihon, Moritz Bohmecke-Schwafert, Kevin Xu, Mirko Boehm, and Knut Blind, 2024, Measuring software innovation with open source software development data, *arXiv e-prints* .

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

Prescott, Edward C, and Michael Visscher, 1980, Organization capital, *Journal of Political Economy* 88, 446–461.

Reeb, David M, and Wanli Zhao, 2020, Patents do not measure innovation success, *Critical Finance Review* 9, 157–199.

Schreieck, Maximilian, Yongli Huang, Alexander Kupfer, and Helmut Krcmar, 2024, The effect of digital platform strategies on firm value in the banking industry, *Journal of Management Information Systems* 41.

Sheoran, Jyoti, Kelly Blincoe, Eirini Kalliamvakou, Daniela Damian, and Jordan Ell, 2014, Understanding "watchers" on github, *MSR 2014: Proceedings of the 11th Working Conference on Mining Software Repositories* 336–339.

West, Joel, and Scott Gallagher, 2006, Challenges of open source innovation: The paradox of firm investment in open-source software, *RD Management* 36, 319–331.

You, Haifeng, and Xiao-jun Zhang, 2009, Financial reporting complexity and investor underreaction to 10-k information, *Review of Accounting Studies* 14, 559–586.

Figure 1. Lag Structure of GitHub-Based Return Predictability

Panel A shows the t-statistics of market equity-scaled PCA GitHub engagement in forecasting stock returns from lag 0 to lag 10 months using excess returns, industry-adjusted returns, FF5 alphas, and HXZ alphas. Panel B shows High-Low spread magnitude and significance counts. Panel C illustrates the return ordering, i.e., whether High > Middle > Low.

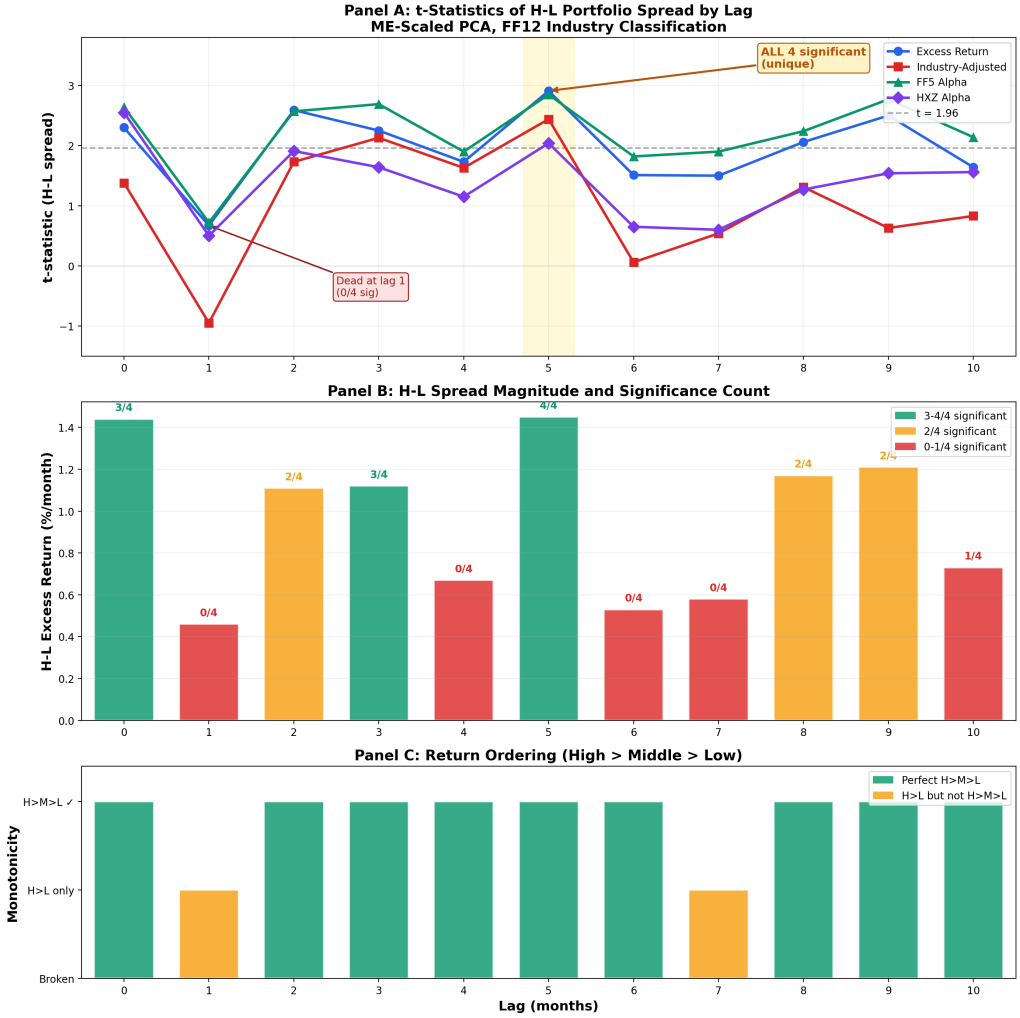

Table 1. Top 10 Companies

This table lists the top ten companies that actively share projects on GitHub and reports the corresponding counts of GitHub engagement.

Company Name	**Ticker**	**Count**
MICROSOFT CORP	MSFT	6,097
INTERNATIONAL BUSINESS MACHINES CORP	IBM	3,131
ALPHABET INC	GOOGL	2,625
INTEL CORP	INTC	1,157
ADOBE INC	ADBE	955
PALO ALTO NETWORKS INC	PANW	419
NVIDIA CORP	NVDA	412
SALESFORCE COM INC	CRM	389
EPAM SYSTEMS INC	EPAM	337
CISCO SYSTEMS INC	CSCO	331

Table 2. Industry Distribution

This table reports annual GitHub engagement counts from 2009 to 2022 across Fama-French 10 industry groups.

Fama-French 10 Industries	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
Consumer Nondurables	0	5	5	3	6	10	11	22	23	31	28	86	36	25
Consumer Durables	0	0	2	4	0	0	0	3	3	1	4	39	7	8
Manufacturing	0	2	1	5	12	17	41	48	49	53	23	31	39	23
Oil, Gas, and Coal Extraction and Products	0	0	0	0	1	0	2	1	10	19	54	27	10	2
Business Equipment	0	1	15	187	461	908	1,474	1,693	2,130	2,657	2,844	3,466	3,048	2,946
Telephone and Television Transmission	0	0	0	1	11	5	30	34	70	94	69	92	46	54
Wholesale, Retail, and Some Services	0	0	0	21	77	72	106	200	111	98	117	96	87	73
Healthcare, Medical Equipment, and Drugs	0	0	0	1	3	4	25	24	24	16	19	28	34	59
Utilities	0	0	0	0	2	3	9	15	2	7	7	3	0	2
Other	0	0	11	38	77	103	199	262	222	327	379	414	356	315

Table 3. Low Middle High Comparison

At the end of June of year t from 2010 to 2022, we sort firms based on the stars received by their GitHub repositories into three groups (Low, Middle, High) based on the 33rd and 67th percentiles of the GitHub stars measure in year t-1. This table reports the time-series mean and median of cross-sectional average characteristics of firms in each GitHub star group. BM ratio is the ratio of book equity of fiscal year ending in year t-1 to market capitalization at the end of year t-1. R&D ratio is R&D expenses in fiscal year ending in year t-1 divided by market capitalization at the end of year t-1. Innovation is ln(1+Patents/Assets), which is the natural log of one plus the number of patents granted in year t-1 divided by total assets in fiscal year ending in year t-1. Skewness of return is computed at the end of June of year t using daily returns over the previous 12 months (with a minimum of 31 trading days).

	Low Mean	**Middle Mean**	**High Mean**	**Low Median**	**Middle Median**	**High Median**
Ln(Total asset)	9.145	9.534	10.473	8.977	9.224	10.645
BM ratio	0.230	0.220	0.188	0.186	0.176	0.151
R&D ratio	0.016	0.018	0.021	0.014	0.016	0.019
Innovation	0.001	0.001	0.001	0.000	0.000	0.000
Skewness of return	-0.023	-0.016	0.010	-0.029	-0.040	0.009
Return volatility	0.020	0.020	0.019	0.017	0.017	0.017
# Firm-month	1,956	1,768	1,674	1,956	1,768	1,674

Table 4. Return Predictive Power of GitHub Activity Measures: Principal Component Analysis Scaled by Market Equity

At the end of every month from 2010 to 2022, we perform principal component analysis (PCA) on four market-equity-scaled measures of GitHub activity: stars, forks, watchers, and total repositories. The principal component is constructed using a rolling-window PCA based solely on information available prior to portfolio formation, thereby avoiding look-ahead bias. Then, from 2010 to 2022, we form portfolios based on the resulting PCA measure. We also construct a high-minus-low (High-Low) portfolio by holding a long (short) position in the high (low) GitHub portfolio. We then hold these portfolios over the next month. In panel A, we report their average monthly returns in excess of the one-month Treasury bill rate (Exret) and their average monthly industry-adjusted returns. The portfolio industry-adjusted returns (Ind-adjret) are calculated as the difference between individual firms' returns and those of firms in the same industry (based on Fama-French 48 industry classifications). In panels B and C, we report the alphas and R^2 from the regression of the time-series of portfolio excess returns on various factor models: the Fama-French five factors (the market factor-MKT, the size factor-SMB, the value factor-HML, the robust-minus-weak factor-RMW, and the conservative-minus-aggressive factor-CMA) as in Fama and French (2015), and alphas from the investment-based factor model (q-factor model) of Hou et al. (2015) (HXZ). All returns and alphas are value-weighted and expressed as a percentage. The t-statistics are computed using Newey and West (1987) standard errors with six lags and are reported below return estimates. R^2 is adjusted.

	Panel A: Excess and adjusted returns		Panel B: Alpha from different factor models		Panel C: R^2 of different factor models	
	Exret	Ind-adjret	FF 5f	HXZ (q factor)	FF 5f	HXZ (q factor)
Low	1.34%	-0.21%	0.21%	0.08%	0.74	0.63
T value	3.45	-1.93	0.89	0.29		
Middle	1.50%	-0.19%	0.25%	0.21%	0.70	0.19
T value	3.28	-1.03	1.07	0.59		
High	2.79%	0.67%	1.75%	1.80%	0.54	0.33
T value	4.74	2.15	3.52	2.31		
High-Low	1.45%	0.88%	1.55%	1.72%	0.01	-0.01
T value	2.91	2.44	2.85	2.04		

Table 5. Long-Horizon Returns of the Market Equity-Scaled PCA GitHub Portfolio

This table displays the High-minus-Low portfolio's average monthly excess returns, average monthly industry-adjusted returns, and alphas and R^2s from the Fama-French five-factor model, and the q-factor model from the beginning of current month to the end of the 12th month, 24th month, 36th month, 48th month, and 60th month.

Year	Exret	Ind-adjret	FF 5f	HXZ q factor	FF 5f R square	HXZ R square
12 months	15.25%	7.78%	15.46%	17.88%	0.01	-0.01
t value	3.84	3.10	3.93	3.53		
24 months	25.71%	11.90%	25.78%	28.71%	-0.03	-0.02
t value	4.46	4.96	4.47	3.82		
36 months	34.78%	16.30%	34.50%	38.16%	-0.01	0.02
t value	5.91	5.71	5.79	5.07		
48 months	42.71%	19.89%	42.56%	49.01%	0.00	0.01
t value	7.30	6.07	7.33	7.42		
60 months	52.48%	22.91%	53.04%	62.46%	-0.01	0.01
t value	9.18	6.04	9.31	11.59		

Table 6. Fama-MacBeth Regression

This table reports the time-series averages of monthly slope coefficients from Fama and MacBeth (1973) cross-sectional regressions over the period 2010-2022. In each month, individual stock returns are regressed on lagged GitHub engagement measures—ln(1+watchers), ln(1+stars), ln(1+forks), and PC1 (the first principal component from a rolling-window PCA of GitHub activity measures)—along with control variables. All accounting controls are measured in fiscal year t-1. Firm and year-month fixed effects are included. Reported t-statistics are based on time-series variation in the monthly slope estimates.

	(1)	(2)	(3)	(4)
L.r0	-0.046***	-0.045***	-0.044***	-0.046***
	(-3.98)	(-3.96)	(-3.92)	(-4.01)
Ln(1+watchers)	0.002***			
	(3.30)			
Ln(1+stars)		0.002***		
		(3.22)		
Ln(1+forks)			0.002***	
			(3.10)	
PC1				0.002***
				(3.37)
Ln(total asset)	0.001*	0.001	0.001*	0.001*
	(1.77)	(1.52)	(1.67)	(1.80)
BM ratio	-0.001	-0.000	-0.001	-0.002
	(-0.51)	(-0.18)	(-0.40)	(-0.61)
R&D ratio	0.038	0.032	0.035	0.036
	(0.67)	(0.57)	(0.63)	(0.65)
Innovation	-0.014	-0.004	-0.021	-0.022
	(-0.25)	(-0.07)	(-0.33)	(-0.35)
Firm FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
N	69,088	69,088	69,088	69,088
Adj. R^2	0.050	0.050	0.051	0.051

Statistical significance is indicated as follows: *** p <0.01, ** p <0.05, * p <0.1.

Table 7. Panel Regression

This table reports results from monthly panel regressions of individual stock returns on lagged GitHub engagement measures over the period 2010-2022. The dependent variable is the monthly stock return. Key explanatory variables include ln(1+watchers), ln(1+stars), ln(1+forks), and PC1 (the first principal component from a rolling-window PCA of GitHub engagement metrics). All regressions include firm and year-month fixed effects. Accounting controls are measured in year t-1. Standard errors are clustered at the firm level.

	(1) r0	(2) r0	(3) r0	(4) r0
L.r0	-0.047***	-0.047***	-0.047***	-0.047***
	(-7.26)	(-7.28)	(-7.29)	(-7.29)
Ln(1+watchers)	0.004***			
	(6.65)			
Ln(1+stars)		0.003***		
		(6.02)		
Ln(1+forks)			0.004***	
			(6.04)	
PC1				0.003***
				(6.44)
Ln(total asset)	-0.006***	-0.006***	-0.006***	-0.006***
	(-4.84)	(-5.08)	(-4.63)	(-4.68)
BM ratio	0.010**	0.011***	0.010**	0.010**
	(2.43)	(2.70)	(2.41)	(2.41)
R&D ratio	-0.089	-0.094	-0.092	-0.091
	(-0.82)	(-0.87)	(-0.85)	(-0.84)
Innovation	0.055	0.064	0.059	0.059
	(0.55)	(0.66)	(0.59)	(0.60)
Firm FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
N	69,088	69,088	69,088	69,088
Adj. R^2	0.016	0.016	0.016	0.016

Statistical significance is indicated as follows: *** p <0.01, ** p <0.05, * p <0.1.

Table 8. Raw GitHub Measures, GitHub PCA Composite, and Innovation Proxies (lag 5)

This table compares the return predictability of GitHub engagement measures—stars, forks, watchers, number of repositories, and their principal component analysis—with conventional innovation proxies (R&D/ME, Patents, and Trademarks). For each metric, firms are sorted monthly into deciles, and long-short portfolios are formed. The table reports equal-weighted and value-weighted excess returns and Newey-West t-statistics with six lags.

Signal	Portfolio	Weighting	Mean Monthly Return (%)	t-stat (NW, 6 lags)	Sharpe (annualized)
Stars (lag 5)	Short	EW	1.62	3.79	1.01
Stars (lag 5)	Long	EW	1.79	4.34	1.12
Stars (lag 5)	Long-Short	EW	0.17	0.52	0.19
Stars (lag 5)	Short	VW	1.30	3.30	0.90
Stars (lag 5)	Long	VW	2.14	5.24	1.34
Stars (lag 5)	Long-Short	VW	0.84	2.27	0.71
Forks (lag 5)	Short	EW	1.62	3.79	1.01
Forks (lag 5)	Long	EW	1.72	4.12	1.08
Forks (lag 5)	Long-Short	EW	0.10	0.35	0.12
Forks (lag 5)	Short	VW	1.31	3.33	0.92
Forks (lag 5)	Long	VW	2.08	5.15	1.32
Forks (lag 5)	Long-Short	VW	0.77	2.18	0.68
Watchers (lag 5)	Short	EW	1.67	3.92	1.05
Watchers (lag 5)	Long	EW	1.80	4.65	1.15
Watchers (lag 5)	Long-Short	EW	0.13	0.49	0.16
Watchers (lag 5)	Short	VW	1.38	3.34	0.92
Watchers (lag 5)	Long	VW	2.08	5.18	1.34
Watchers (lag 5)	Long-Short	VW	0.69	2.21	0.65
Repositories (lag 5)	Short	EW	1.62	3.79	1.01
Repositories (lag 5)	Long	EW	1.79	4.73	1.18
Repositories (lag 5)	Long-Short	EW	0.17	0.61	0.20
Repositories (lag 5)	Short	VW	1.30	3.30	0.90
Repositories (lag 5)	Long	VW	2.02	5.22	1.33
Repositories (lag 5)	Long-Short	VW	0.72	1.94	0.63
GitHub PCA (raw, lag 5)	Short	EW	1.62	3.79	1.01
GitHub PCA (raw, lag 5)	Long	EW	1.77	4.50	1.17
GitHub PCA (raw, lag 5)	Long-Short	EW	0.15	0.49	0.17
GitHub PCA (raw, lag 5)	Short	VW	1.30	3.30	0.90
GitHub PCA (raw, lag 5)	Long	VW	2.07	5.15	1.34
GitHub PCA (raw, lag 5)	Long-Short	VW	0.77	2.08	0.67
R&D/ME (lag 5)	Short	EW	1.10	2.59	0.67
R&D/ME (lag 5)	Long	EW	1.87	4.43	1.07
R&D/ME (lag 5)	Long-Short	EW	0.77	2.41	0.66
R&D/ME (lag 5)	Short	VW	1.37	4.36	0.98
R&D/ME (lag 5)	Long	VW	2.48	4.69	1.31
R&D/ME (lag 5)	Long-Short	VW	1.11	2.27	0.74
Patents (lag 5)	Short	EW	1.39	3.23	0.84
Patents (lag 5)	Long	EW	1.62	3.99	1.05
Patents (lag 5)	Long-Short	EW	0.23	0.90	0.27
Patents (lag 5)	Short	VW	1.74	4.34	1.19
Patents (lag 5)	Long	VW	2.06	5.02	1.17
Patents (lag 5)	Long-Short	VW	0.31	0.89	0.26
Trademarks (lag 5)	Short	EW	1.16	3.12	0.81
Trademarks (lag 5)	Long	EW	1.55	3.71	0.95
Trademarks (lag 5)	Long-Short	EW	0.39	1.41	0.44
Trademarks (lag 5)	Short	VW	1.65	4.74	1.19
Trademarks (lag 5)	Long	VW	2.17	4.31	1.12
Trademarks (lag 5)	Long-Short	VW	0.53	1.09	0.37

Table 9. Out-of-Sample Predictive Performance of GitHub Engagement Measures vs. Traditional Innovation Proxies

This table reports out-of-sample R^2_{OOS} values from forecasting monthly stock returns using GitHub-based engagement measures—OSAI and OSAI Peer-Adjusted—alongside conventional innovation proxies, including R&D/ME, patents, and trademarks. For each signal, we construct predictive regressions using either the overall sample or subsamples formed by size terciles, and evaluate performance using one-month-ahead or five-month-ahead forecasts. Higher R^2_{OOS} indicates better out-of-sample predictive accuracy.

Signal	Bucket	R^2_{OOS}
OSAI (lag 1)	Overall	-0.00037
OSAI (lag 1)	Size Tercile 1	-0.00033
OSAI (lag 1)	Size Tercile 2	-0.00051
OSAI (lag 1)	Size Tercile 3	-0.00024
OSAI Peer-Adj (lag 1)	Overall	-0.00041
OSAI Peer-Adj (lag 1)	Size Tercile 1	-0.00031
OSAI Peer-Adj (lag 1)	Size Tercile 2	-0.00053
OSAI Peer-Adj (lag 1)	Size Tercile 3	-0.00042
R&D/ME z (lag 1)	Overall	0.00001
R&D/ME z (lag 1)	Size Tercile 1	-0.00166
R&D/ME z (lag 1)	Size Tercile 2	0.00143
R&D/ME z (lag 1)	Size Tercile 3	0.00106
Patents (lag 1)	Overall	0.00024
Patents (lag 1)	Size Tercile 1	0.00124
Patents (lag 1)	Size Tercile 2	0.00018
Patents (lag 1)	Size Tercile 3	-0.00145
Trademarks (lag 1)	Overall	-0.00067
Trademarks (lag 1)	Size Tercile 1	-0.00081
Trademarks (lag 1)	Size Tercile 2	-0.00022
Trademarks (lag 1)	Size Tercile 3	-0.00102
OSAI (lag 5)	Overall	0.00013
OSAI (lag 5)	Size Tercile 1	0.00124
OSAI (lag 5)	Size Tercile 2	-0.00091
OSAI (lag 5)	Size Tercile 3	-0.00042
OSAI Peer-Adj (lag 5)	Overall	-0.00027
OSAI Peer-Adj (lag 5)	Size Tercile 1	0.0009
OSAI Peer-Adj (lag 5)	Size Tercile 2	-0.00161
OSAI Peer-Adj (lag 5)	Size Tercile 3	-0.00055

Appendix

Table A1. Return Predictive Power of GitHub Activity Measures: Excluding Microsoft

This table repeats the analysis in Table 4 while excluding Microsoft. In panel A, we report portfolio average monthly returns in excess of one-month Treasury bill rate (Exret) and their average monthly industry-adjusted returns. In panels B and C, we report the alphas and R^2 from the regression of the time-series of portfolio excess returns on various factor models: the Fama-French five factors as in Fama and French (2015), and alphas from the investment-based factor model (q-factor model) of Hou et al. (2015) (HXZ). All returns and alphas are value-weighted and expressed in percentage. The t-statistics are computed using Newey and West (1987) standard errors with six lags and are reported below return estimates. R^2 is adjusted.

	Panel A: Excess and adjusted returns		Panel B: Alpha from different factor models		Panel C: R^2 of different factor models	
	Exret	Ind-adjret	FF 5f	HXZ (q factor)	FF 5f	HXZ (q factor)
Low	1.29%	-0.25%	0.14%	-0.02%	0.74	0.63
T value	3.37	-2.88	0.63	-0.07		
Middle	1.48%	-0.09%	0.28%	0.39%	0.72	0.19
T value	3.15	-0.49	1.34	1.62		
High	2.70%	0.66%	1.63%	1.62%	0.54	0.31
T value	4.02	1.73	2.98	1.96		
High-Low	1.41%	0.90%	1.49%	1.64%	-0.01	-0.02
T value	2.66	2.15	2.69	1.92		

Table A2. Return Predictive Power of the Number of Stars Received by GitHub Repositories after Deleting Fake Stars: Single-Sorted Portfolio Analysis

At the end of every month from 2010 to 2022, we form portfolios based on the number of stars received by firms' GitHub repositories after deleting fake stars. Fake stars have been reported by He et al. (2024) wherein bots have artificially inflated the number of stars in various repositories. We also construct a high-minus-low (High-Low) portfolio by holding a long (short) position in the high (low) GitHub portfolio. We then hold these portfolios over the next month. In panel A, we report their average monthly returns in excess of one-month Treasury bill rate (Exret) and their average monthly industry-adjusted returns. The portfolio industry-adjusted returns (Ind-adjret) are based on the difference between individual firms' returns and the returns of firms in the same industry (based on Fama-French 48 industry classifications). In panels B and C, we report the alphas and R^2 from the regression of the time-series of portfolio excess returns on various factor models: the Fama-French five factors (the market factor-MKT, the size factor-SMB, the value factor-HML, the robust-minus-weak factor-RMW, and the conservative-minus-aggressive factor-CMA) as in Fama and French (2015), and alphas from the investment-based factor model (q-factor model) of Hou et al. (2015) (HXZ). All returns and alphas are value-weighted and expressed in percentage. The t-statistics are computed using Newey and West (1987) standard errors with six lags and are reported below return estimates. R^2 is adjusted.

	Panel A: Excess and adjusted returns		Panel B: Alpha from different factor models		Panel C: R^2 of different factor models	
	Exret	Ind-adjret	FF 5f	HXZ (q factor)	FF 5f	HXZ (q factor)
Low	0.77%	-0.22%	-0.38%	-0.67%	0.66	0.57
T value	1.92	-1.10	-1.47	-1.91		
Middle	1.04%	-0.17%	0.06%	0.38%	0.59	0.18
T value	3.21	-1.16	0.27	1.22		
High	1.92%	0.27%	0.76%	0.84%	0.70	0.47
T value	4.54	2.62	3.02	2.23		
High-Low	1.16%	0.49%	1.14%	1.52%	0.07	0.01
T value	3.25	2.09	3.26	3.18		

Table A3. Return Predictive Power of Hard-to-Fake Metrics: Single-Sorted Portfolio Analysis

This table summarizes high-minus-low (High-Low) portfolio returns and alphas for hard-to-fake metrics. Hard-to-fake metrics include forks, pull requests (PRs), contributors, issues, and commits. At the end of every month from 2010 to 2022, we form portfolios based on the corresponding hard-to-fake metric. We construct a High-Low portfolio by holding a long (short) position in the high (low) GitHub portfolio. We then hold these portfolios over the next month. In panel A, we report their average monthly returns in excess of one-month Treasury bill rate (Exret) and their average monthly industry-adjusted returns. The portfolio industry-adjusted returns (Ind-adjret) are based on the difference between individual firms' returns and the returns of firms in the same industry (based on Fama-French 48 industry classifications). In panels B and C, we report the alphas and R^2 from the regression of the time-series of portfolio excess returns on various factor models: the Fama-French five factors (the market factor-MKT, the size factor-SMB, the value factor-HML, the robust-minus-weak factor-RMW, and the conservative-minus-aggressive factor-CMA) as in Fama and French (2015), and alphas from the investment-based factor model (q-factor model) of Hou et al. (2015) (HXZ). All returns and alphas are value-weighted and expressed in percentage. The t-statistics are computed using Newey and West (1987) standard errors with six lags and are reported below return estimates. R^2 is adjusted.

		Panel A: Excess and adjusted returns		Panel B: Alpha from different factor models		Panel C: R^2 of different factor models	
Metric		Exret	Ind-adjret	FF 5f	HXZ (q factor)	FF 5f	HXZ (q factor)
Forks	High-Low	1.18%	0.35%	1.12%	1.11%	0.11	0.08
	T value	3.18	1.66	3.04	2.92		
PRs	High-Low	1.14%	0.31%	1.07%	1.05%	0.04	0.03
	T value	3.67	1.52	3.37	3.24		
Contributors	High-Low	1.04%	0.37%	0.96%	2.14%	0.03	0.02
	T value	3.27	1.67	2.99	2.87		
Issues	High-Low	1.07%	0.39%	1.00%	0.98%	0.03	0.02
	T value	3.30	1.72	3.00	2.88		
Commits	High-Low	1.07%	0.40%	1.02%	1.00%	0.03	0.02
	T value	3.35	1.77	3.09	2.97		

Table A4. GitHub Sharing Mimicking Portfolio and Risk Premium

This table reports regression results for a GitHub-based long-minus-short (High-Low) portfolio constructed by sorting firms into terciles based on GitHub repository sharing and taking a long position in the top tercile and a short position in the bottom tercile. The table presents the results from the second-pass cross-sectional regressions of average portfolio returns on factor betas. Test assets include single-sorted portfolios from Hou et al. (2020) in Panel A and long-minus-short portfolios from Hou et al. (2020) in Panel B. "Sharing L-S" is the GitHub sharing mimicking portfolio; "mktrf, smb, hml, rmw, cma, umd" are Fama and French (2018)'s six factors; "r_mkt, r_me, r_ia, r_roe, r_eg" are Hou et al. (2021)'s Q5 factors; and "pead" and "fin" are Daniel et al. (2020)'s behavioral factors. Reported are monthly risk premium, R^2 in percentages, and p-values. N is the number of test portfolios and T is the number of months.

Panel A: Single-Sorted Portfolios from Hou et al. (2020)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	102.343**	136.484**	92.318***	103.992***	79.968**	96.049***	110.279***	84.027*
	(0.019)	(0.019)	(0.007)	(0.004)	(0.044)	(0.008)	(0.004)	(0.053)
Sharing L-S	1.344	2.032**				1.76**	1.999***	2.051**
	(0.389)	(0.016)				(0.014)	(0.006)	(0.015)
mktrf		-0.279	0.137	0.042	0.295	0.111	-0.008	0.263
		(0.689)	(0.785)	(0.935)	(0.589)	(0.831)	(0.988)	(0.646)
smb			-0.076			-0.122		
			(0.797)			(0.685)		
hml			-0.400			-0.283		
			(0.276)			(0.450)		
rmw			-0.114			-0.088		
			(0.607)			(0.698)		
cma			0.113			0.089		
			(0.625)			(0.707)		
umd			0.174			0.091		
			(0.623)			(0.798)		
r_mkt				3.967			-1.021	
				(0.938)			(0.985)	
r_me				-4.811			-14.444	
				(0.881)			(0.666)	
r_ia				5.963			6.290	
				(0.826)			(0.821)	
r_roe				-7.150			-4.846	
				(0.798)			(0.865)	
r_eg				31.158			28.333	
				(0.328)			(0.391)	
pead					34.743			32.937
					(0.215)			(0.264)
fin					-7.000			-3.691
					(0.855)			(0.925)
R^2	6.035	13.306	18.081	15.156	15.716	26.224	24.755	27.910
N	397	397	397	397	397	397	397	397
T	134	134	134	134	134	134	134	134

Statistical significance is indicated as follows: *** p <0.01, ** p <0.05, * p <0.1.

Panel B: 118 Long-Minus-Short Portfolios from Hou et al. (2020)								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Intercept	3.672	-16.057	3.079	-13.347	-0.942	-3.713	-17.389	-6.749
	(0.160)	(0.422)	(0.809)	(0.330)	(0.938)	(0.799)	(0.352)	(0.695)
Sharing L-S		5.208*				3.576*	4.905*	4.571*
		(0.057)				(0.096)	(0.060)	(0.055)
mktrf		-0.456	-0.997	-0.535	-0.244	-1.280	-0.988	-0.692
		(0.632)	(0.197)	(0.508)	(0.749)	(0.164)	(0.360)	(0.479)
smb			0.342			0.284		
			(0.475)			(0.604)		
hml			-1.087**			-0.674		
			(0.046)			(0.292)		
rmw			-0.359			-0.414		
			(0.341)			(0.352)		
cma			0.085			0.088		
			(0.782)			(0.800)		
umd			0.458			0.190		
			(0.296)			(0.688)		
r_mkt				-53.817			-99.104	
				(0.504)			(0.357)	
r_me				-24.688			-19.225	
				(0.620)			(0.765)	
r_ia				3.599			3.491	
				(0.920)			(0.937)	
r_roe				-40.785			-38.169	
				(0.255)			(0.380)	
r_eg				41.869			19.621	
				(0.289)			(0.678)	
pead					39.182			10.395
					(0.207)			(0.791)
fin					-54.764			-67.757
					(0.279)			(0.286)
R^2	12.943	25.336	35.624	21.993	14.175	46.279	42.537	30.660
N	42	42	42	42	42	42	42	42
T	134	134	134	134	134	134	134	134

Statistical significance is indicated as follows: *** p <0.01, ** p <0.05, * p <0.1.

Table A5. GitHub Stars Decomposition

This table reports the results from decomposition of cross-sectional variation in the number of stars a firm's GitHub repository has received. In the spirit of Cho et al. (2024), we compute the discounted sums of future ROE, future log returns, and future IVA based on realizations over a multiyear horizon. We form 3 GitHub stars portfolios and then compute the relevant variables at the portfolio level for the subsequent J years and run cross-sectional regressions of each of the components on the number of GitHub stars.

Time Horizon	ROE Contribution Coefficient	ER Contribution Coefficient	IVA Contribution Coefficient
2 year	0.2541	0.1575	0.1022
4 year	0.3165	0.1610	0.1066
6 year	0.3238	0.1630	0.1070
Infinity	0.3281	0.1650	0.1081

Table A6. Return Predictive Power of GitHub-Based Signals and Traditional Innovation Proxies: Single-Sorted Portfolio Analysis—Lag 1

This table compares the return predictability of GitHub engagement measures—OSAI and OSAI Peer-Adjusted—with conventional innovation proxies (R&D/ME, Patents, and Trademarks). OSAI is a composite, size-normalized GitHub activity index constructed from residualized and standardized $ln(1+x)$ counts of forks, pull requests, contributors, issues, and commits. OSAI Peer-Adjusted further demeans OSAI within size x R&D-intensity peer groups. For each metric, firms are sorted monthly into deciles, and long-short portfolios are formed. The table reports equal-weighted and value-weighted excess returns, Newey-West t-statistics with six lags, and alphas from the Fama-French five-factor model.

Signal	(1) Portfolio	(2) Weighting	(3) Mean Monthly Return	(4) t-stat (NW, 6 lags)	(5) Sharpe (annualized)	(6) Alpha (monthly)
OSAI (lag 1)	Short	EW	1.86	4.44	0.92	
OSAI (lag 1)	Long	EW	1.32	1.85	0.65	
OSAI (lag 1)	Long-Short	EW	-0.54	-0.69	-0.23	-0.69
OSAI (lag 1)	Short	VW	1.69	3.33	0.77	
OSAI (lag 1)	Long	VW	1.14	1.74	0.57	
OSAI (lag 1)	Long-Short	VW	-0.55	-0.72	-0.23	-0.59
OSAI Peer-Adj (sizexR&D, lag 1)	Short	EW	1.11	2.46	0.54	
OSAI Peer-Adj (sizexR&D, lag 1)	Long	EW	1.42	2.07	0.71	
OSAI Peer-Adj (sizexR&D, lag 1)	Long-Short	EW	0.31	0.49	0.15	0.26
OSAI Peer-Adj (sizexR&D, lag 1)	Short	VW	0.92	1.77	0.43	
OSAI Peer-Adj (sizexR&D, lag 1)	Long	VW	1.22	1.89	0.62	
OSAI Peer-Adj (sizexR&D, lag 1)	Long-Short	VW	0.30	0.50	0.15	0.37
R&D/ME z (lag 1)	Short	EW	0.05	0.09	0.02	
R&D/ME z (lag 1)	Long	EW	1.69	2.72	0.72	
R&D/ME z (lag 1)	Long-Short	EW	1.64	2.46	0.74	1.60
R&D/ME z (lag 1)	Short	VW	0.41	0.71	0.22	
R&D/ME z (lag 1)	Long	VW	2.07	2.59	0.75	
R&D/ME z (lag 1)	Long-Short	VW	1.66	2.30	0.66	1.72
Patents (lag 1)	Short	EW	0.77	1.26	0.40	
Patents (lag 1)	Long	EW	1.74	3.52	0.85	
Patents (lag 1)	Long-Short	EW	0.97	1.37	0.49	1.07
Patents (lag 1)	Short	VW	0.73	1.13	0.39	
Patents (lag 1)	Long	VW	2.14	3.52	0.96	
Patents (lag 1)	Long-Short	VW	1.40	1.92	0.70	1.49
Trademarks (lag 1)	Short	EW	1.10	2.22	0.52	
Trademarks (lag 1)	Long	EW	1.93	2.69	0.66	
Trademarks (lag 1)	Long-Short	EW	0.83	1.04	0.29	0.60
Trademarks (lag 1)	Short	VW	1.37	2.73	0.63	
Trademarks (lag 1)	Long	VW	2.18	2.56	0.69	
Trademarks (lag 1)	Long-Short	VW	0.81	0.81	0.27	0.86

Table A7. Out-of-Sample Predictive Performance: Deep Learning Robustness Test

This table presents out-of-sample performance of a long short-term memory (LSTM) model with and without GitHub engagement. The augmented model incorporates firm-level GitHub star counts in addition to standard Fama-French factors and is evaluated using an expanding-window forecasting design.

Model Specification	GitHub Signal Included	Test MSE	R^2_{OOS}	ΔR^2_{OOS} vs. Base
Base Model	No	0.7055	0.2898	—
Enhanced Model	Stars	0.7008	0.2946	+0.0048

Table A8. Difference-in-Differences Analysis Using GitHub's Plan Changes

This table presents the results of a difference-in-differences analysis around GitHub's plan changes on April 14, 2020. On that date, GitHub announced that all of its core features are available for free to all users including those on free accounts. In column (1), the variable of interest is Post * treat_watchers. Post is 1 after April 2020, 0 prior to that date. Treat_watchers is 1 if the firm's number of watchers is larger than the median number of watchers in the 12 months before April 2020. Columns (2) to (4) replace watchers with stars, forks, and repository count, respectively. In all specifications, the lagged value of r0 is controlled for.

	(1) r0	(2) r0	(3) r0	(4) r0
L.r0	-0.060***	-0.060***	-0.060***	-0.060***
	(-15.80)	(-15.80)	(-15.80)	(-15.79)
Post * treat_watchers	0.003*			
	(1.81)			
Post * treat_stars		0.003**		
		(1.99)		
Post * treat_forks			0.003**	
			(2.05)	
Post * treat_count				0.002*
				(1.80)
Ln(total asset)	-0.009***	-0.009***	-0.009***	-0.009***
	(-11.05)	(-11.06)	(-11.07)	(-11.06)
BM ratio	0.024***	0.024***	0.024***	0.024***
	(10.53)	(10.56)	(10.56)	(10.54)
R&D ratio	-0.113**	-0.113**	-0.113**	-0.113**
	(-2.31)	(-2.32)	(-2.32)	(-2.32)
Innovation	0.182***	0.183***	0.183***	0.183***
	(4.64)	(4.67)	(4.66)	(4.65)
Firm FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
N	69,047	69,047	69,047	69,047
Adj. R^2	0.016	0.016	0.016	0.016

Statistical significance is indicated as follows: *** p <0.01, ** p <0.05, * p <0.1.

Table A9. Parallel Trends Assumption Test

Difference-in-differences estimates showing monthly stock return effects of GitHub engagement before and after the GitHub's plan changes in April 2020.

	(1) Post * treat_watchers r0	(2) Post * treat_stars r0	(3) Post * treat_forks r0	(4) Post * treat_count r0
L.r0	-0.060***	-0.060***	-0.060***	-0.060***
	(-15.79)	(-15.79)	(-15.79)	(-15.78)
DID_-10	-0.002	0.004	0.006	0.008
	(-0.34)	(0.56)	(0.80)	(1.14)
DID_-9	0.004	0.002	0.000	0.005
	(0.52)	(0.33)	(0.04)	(0.70)
DID_-8	-0.007	-0.006	-0.003	-0.004
	(-1.01)	(-0.85)	(-0.49)	(-0.66)
DID_-7	0.004	0.007	0.008	0.007
	(0.54)	(1.00)	(1.14)	(1.04)
DID_-6	0.004	0.004	0.004	0.004
	(0.51)	(0.64)	(0.55)	(0.53)
DID_-5	0.003	0.002	0.001	0.006
	(0.45)	(0.22)	(0.11)	(0.84)
DID_-4	0.010	0.013*	0.011	0.011*
	(1.46)	(1.93)	(1.61)	(1.70)
DID_-3	0.001	0.002	0.003	-0.001
	(0.11)	(0.34)	(0.44)	(-0.19)
DID_-2	0.004	0.006	0.005	0.006
	(0.60)	(0.92)	(0.76)	(0.90)
DID_-1	0.001	0.004	0.006	0.003
	(0.19)	(0.79)	(1.17)	(0.63)
DID_1	0.012*	0.009	0.009	0.009
	(1.76)	(1.34)	(1.26)	(1.39)
DID_2	-0.004	-0.003	-0.009	-0.005
	(-0.60)	(-0.47)	(-1.27)	(-0.70)
DID_3	0.006	0.006	0.011	0.007
	(0.82)	(0.86)	(1.63)	(0.98)
DID_4	0.025***	0.020***	0.016**	0.018***
	(3.54)	(2.90)	(2.30)	(2.67)
DID_5	-0.010	-0.009	-0.010	-0.004
	(-1.43)	(-1.27)	(-1.36)	(-0.62)
DID_6	-0.007	-0.005	-0.005	-0.002
	(-1.03)	(-0.78)	(-0.67)	(-0.25)
DID_7	0.010	0.009	0.007	0.005
	(1.44)	(1.36)	(1.05)	(0.69)
DID_8	0.015**	0.014**	0.013*	0.015**
	(2.10)	(1.96)	(1.82)	(2.25)
DID_9	0.014*	0.016**	0.016**	0.017**
	(1.96)	(2.35)	(2.22)	(2.53)
DID_10	0.006	0.001	0.003	0.001
	(0.81)	(0.18)	(0.39)	(0.20)
Ln(total asset)	-0.009***	-0.009***	-0.009***	-0.009***
	(-11.05)	(-11.05)	(-11.05)	(-11.04)
BM ratio	0.024***	0.024***	0.024***	0.024***
	(10.57)	(10.60)	(10.59)	(10.61)
R&D ratio	-0.110**	-0.112**	-0.112**	-0.112**
	(-2.26)	(-2.29)	(-2.29)	(-2.29)
Innovation	0.179***	0.180***	0.179***	0.180***
	(4.57)	(4.59)	(4.57)	(4.60)
Firm FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
N	69,047	69,047	69,047	69,047
Adj. R^2	0.016	0.016	0.016	0.016

Statistical significance is indicated as follows: *** p <0.01, ** p <0.05, * p <0.1.

Table A10. Heterogeneous Treatment Effects of GitHub Engagement Intensity: Difference-in-Differences Estimates

This table presents difference-in-differences estimates where the treatment effect is interacted with measures of GitHub engagement intensity, including watchers per employee, stars per employee, forks per employee, and repository count per employee. Coefficients on the interaction terms capture how the treatment effect varies with engagement intensity.

	(1) r0	(2) r0	(3) r0	(4) r0
L.r0	-0.060***	-0.060***	-0.060***	-0.060***
	(-15.69)	(-15.70)	(-15.69)	(-15.69)
watchers per employee	0.000			
	(1.37)			
DID * watchers per employee	-0.000			
	(-1.07)			
stars per employee		0.000***		
		(5.16)		
DID * stars per employee		0.000**		
		(2.50)		
forks per employee			0.000***	
			(4.09)	
DID * forks per employee			0.002***	
			(4.38)	
count per employee				-0.009**
				(-2.56)
DID * count per employee				0.020***
				(3.95)
Ln(total asset)	-0.009***	-0.009***	-0.009***	-0.009***
	(-10.98)	(-11.06)	(-11.10)	(-11.16)
BM ratio	0.024***	0.024***	0.024***	0.024***
	(10.29)	(10.36)	(10.43)	(10.41)
R&D ratio	-0.112**	-0.109**	-0.112**	-0.113**
	(-2.29)	(-2.24)	(-2.30)	(-2.32)
Innovation	0.180***	0.185***	0.188***	0.180***
	(4.58)	(4.71)	(4.78)	(4.59)
Firm FE	Yes	Yes	Yes	Yes
Year-month FE	Yes	Yes	Yes	Yes
N	68,961	68,961	68,961	68,961
Adj. R^2	0.016	0.016	0.016	0.016

Statistical significance is indicated as follows: *** p <0.01, ** p <0.05, * p <0.1.