The Private Value of Open-Source Innovation*

Logan P. Emery[†] **Chan Lim**[‡] **Shiwei Ye**[†]

September 2025

Abstract

Open-source innovation lacks the legal excludability viewed as essential for generating private value from innovation. Nonetheless, using investor reactions to GitHub releases by U.S. public firms from 2015-2023, we estimate an average private value of $849,000 per project. Extrapolation to all projects during this period implies a total value of $909 billion. Firms facing less competition release more projects, and both lower competition and restrictive licenses generate more private value. This value predicts firm growth, but peer benefits are modest. Overall, firms generate private value from open-source innovation by limiting spillovers, challenging the notion that open source fosters industry-wide growth.

Keywords: Open Source, Innovation, Firm growth, Valuation, GitHub
JEL Classification: G14, G30, L21, O36

*We thank Goeun Choi, Fabrizio Core, Mircea Epure, Fabian Gaessler, Shan Huang, Xing Liu, Pengfei Ma, Mikael Paaso, Anthony Rice, Ekaterina Volkova, Michael Woeppel, and seminar participants at the Midwest Finance Association Annual Meeting 2025, ABFER Annual Meeting 2025, Joint Conference with the Allied Korea Finance Associations 2025, SAIF Annual Research Conference 2025, University of Barcelona Micro Workshop 2025, CICF Annual Meeting 2025, ENTFIN Conference Annual Meeting 2025, European Finance Association Annual Meeting 2025, Future Finance Fest 2025, Erasmus University Rotterdam, Florida State University, Leibniz Institute SAFE, Monash University, Rochester Institute of Technology, Sungkyunkwan University, and University at Buffalo for helpful comments. We are responsible for any remaining errors. Emails: emery@rsm.nl, chan.lim@pepperdine.edu, ye@rsm.nl.
[†]Rotterdam School of Management, Erasmus University Rotterdam.
[‡]Seaver College, Pepperdine University.

1 Introduction

Innovation has long been recognized as essential for economic growth (Schumpeter, 1912). In recent years, growth from innovation has increasingly come from software-centric fields such as artificial intelligence and cloud computing (Ahmadi et al., 2024). While innovation is traditionally measured by outputs such as patents, in these fields, significant innovations are often the result of incremental and fast-paced contributions that are difficult to patent (Hall and MacGarvie, 2010). Thus, developers are increasingly turning to a different system of innovation: open source.[1] When an innovation is "open-sourced," it is made publicly available to all parties at little or no cost. A recent survey finds that 90% of Fortune 100 companies use GitHub, the largest platform for developing open-source innovation.[2] However, it remains unclear what value these profit-maximizing entities derive from making their innovation freely available. Nonetheless, assessing the value of open-source innovation has become essential for a complete picture of how innovation contributes to economic growth.

In this paper, we study the private value generated from open-source innovation by publicly traded firms. Conceptually, whether open-source innovation generates any private value is not immediately clear. Conventional innovation systems (e.g., patents) generate private value by granting inventors exclusive rights to monetize their innovation for a given period of time. This legal excludability is seen as crucial for deriving private value from innovation (Arrow, 1962; Crouzet et al., 2022). However, the open-source system lacks this legal excludability, which can result in positive spillovers to competitors. These spillovers can help firms by facilitating iterative technological development, but they can also hurt firms by improving competitors' products. Prior studies have also suggested firms can indirectly generate private value from open-source innovation through benefits from increased adop-

[1] For example, TensorFlow, an open-source machine learning library developed by Google, has had a significant impact on the growth of AI technology, enabling researchers and developers to build and deploy machine learning models without any associated patents.

[2] See https://octoverse.github.com/2022/.

tion, the open-source community, or competitive effects.[3] Research on estimating the value of open-source innovation has focused on using cost-to-replicate approaches, although this may not fully capture indirect channels of value generation (Greenstein and Nagle, 2014; Murciano-Goroff et al., 2021; Robbins et al., 2021; Blind et al., 2021). Thus, existing approaches present a gap in understanding the value created by open-source innovation and how it is distributed across firms and their competitors.

Our analysis proceeds in three parts. First, we document the extent to which publicly traded firms produce open-source innovation and characterize the type of firm that chooses to develop their innovation via open source. Second, we leverage financial markets to estimate the private value generated by open-source innovation and investigate which characteristics of innovation, firms, and product markets are most strongly correlated with private value. Finally, we examine the relation between open-source innovation and future growth for both the firm and its peers.

Our analysis is based on public-firm activity on GitHub. While not all open-source innovation takes place on GitHub, it is the largest platform for developing open-source innovation, specifically computer software, and has become synonymous with the idea of open source. We compile a comprehensive dataset of public-firm activity on GitHub from 2015 through 2023. While only 18% of public firms share open-source innovations on GitHub, these firms represent 68% of the total stock market capitalization and 80% of the total research and development expenditure by public firms in 2023. Moreover, while 32% of these firms are from the "Computer Software" industry, 86% of industries have at least one such firm,[4] demonstrating the growing scope of both software and open-source innovation. Firms participating in open-source innovation are larger, more valuable, and more innovative on average. Furthermore, firms facing less competition are more likely to participate in open source, consistent with concerns about spillovers to competitors. In a regression setting,

[3] See Internet Appendix A for an overview of this literature.

[4] This analysis is based on the Fama-French 49 industries classification system.

most of these differences are absorbed by firm fixed effects, suggesting that firm fixed effects can account for much of the selection bias in open-source activity.

We next employ a modified version of the method developed by Kogan et al. (2017) to measure the private value of open-source innovations on GitHub (which are called "repositories"), as estimated by investors. The methodology relies on firm-specific stock returns over the three days following a repository's release. The resulting estimates provide an ex-ante measure of the economic value captured by the firm (i.e., private value), excluding any value generated for others (i.e., public value), reflecting both the value of the innovation and the value of being open source. The average repository in our sample generates $849,443 (in 2023 dollars), with average values increasing significantly over time.[5] Repository value is highly skewed: the most valuable GitHub portfolios are owned by Amazon.com Inc. and Microsoft Corp., both exceeding $7 billion. Repositories written in Python and those related to AI and its applications are the most valuable. While our estimates cover only a subset of repositories released by public U.S. firms, we provide a back-of-the-envelope calculation of the total value (private and peer value, discussed below) generated by all repositories released on GitHub from 2015 to 2023, which totals approximately $909 billion.

To validate our estimation strategy, we investigate several important issues related to attributing stock returns to repository releases. First, evidence from anecdotes of repository releases, earnings-call transcripts, and trading volume around releases indicates that investors are aware of repository releases. Second, we find limited evidence for strategically timed repository releases, which induces, if anything, an underestimation of value for the affected repositories. Third, concerns about investor inattentiveness, strategic timing, information leakage, or releases conveying other signals (e.g., changes in innovation strategy) hypothesize that investor reactions may be uninformative about repository value. Contrary

[5] In comparison, average patent values are $53 million over the same period, per data provided by Kogan et al. (2017). Thus, the legal excludability enforced by the patent system appears highly valuable. However, the types of innovation contained in patents often differ significantly from those shared on open-source platforms. Anecdotally, developers more often decide between open-sourcing their software or classifying it as a trade secret, especially following recent rulings regarding the patentability of software (Acikalin et al., 2025). This selection in the sample of open-source innovation is an important consideration for our analysis.

to these concerns, we find a significant relation between investor reactions and future repository popularity. Placebo tests using randomly generated release dates do not replicate this relation, further supporting the conclusion that investor reactions contain value-relevant information.

When comparing the open-source system with conventional systems, two sources of value stand out. Conventional systems generate private value through exclusive rights, whereas open-source systems, lacking such legal exclusivity, can create private value by limiting spillovers (e.g., using "restrictive" licenses that limit commercial use of the repository) or by complementing existing products. To assess these channels, we examine the determinants of private value at the repository level. We find that repositories with restrictive licenses are more valuable on average than repositories with fully permissive licenses. This result highlights the value of excludability even in an open source setting and reflects concerns regarding spillovers to competitors. We also find evidence against complementarity between open-source projects and commercial products being a first-order driver of value. Instead, standalone open-source projects tend to be more valuable and account for a larger fraction of firms' open-source-portfolio value on average. Additionally, larger repositories (e.g., more lines of code) are not necessarily more valuable, and repositories with more subsequent issues opened (e.g., bugs) are viewed as less valuable at the time of release.

Because open-source systems are expected to generate more spillovers than conventional systems, we also investigate how product market characteristics correlate with the private value of open-source innovation. We find that firms facing less competition tend to release repositories that generate more private value. This may reflect their capability to capture a larger share of the total value created (i.e., private plus public value). Alternatively, these firms may be more willing to share valuable innovation due to fewer concerns about spillovers to competitors. In either case, our findings highlight competition as a key factor in the private value of open-source innovation. Controlling for competition, firms that are more likely to benefit from spillover effects in the product market produce more-valuable repositories, which

is consistent with the importance of network effects for open-source value.

The preceding results show how licensing choices and product-market competition, both closely related to the potential for positive spillovers, shape the private value of open-source innovation. To complement this, we next consider how managers themselves discuss the private value of open source. Specifically, we analyze earnings calls using a large language model and a list of 17 potential sources of private value compiled from the literature. The results provide corroborating evidence for many aspects of our earlier findings, but also uncover additional nuance. First, sources related to increased adoption are most frequently discussed, followed by sources related to competition concerns. Second, the most discussed source of private value is complementarity with commercial products, revealing a disconnect between how managers advertise their open-source activities and what investors value. Finally, growing the market at-large was the most discussed source of value in the mid-2000s but has since declined; this trend is consistent with firms becoming increasingly concerned with the competitive risks of open-source engagement.

Finally, we investigate the relation between open-source innovation and firms' future growth. Controlling for patented innovation, firms that generate more private value from their open-source projects have a larger growth in sales, profits, employment, and both the number and value of patents granted over the next three years. These results suggest that open-source innovation contributes to firms' long-term growth while also complementing their traditional patent-based innovative capacity. In contrast, the private value generated by competitors' open-source innovation negatively predicts a firm's growth, which is a typical sign of creative destruction. Notably, this effect is concentrated in repositories with restrictive licenses. While we do not stress a causal interpretation of these results, the economic magnitude is comparable to that of patented innovation, challenging the notion that competitors can benefit from open-source innovation.

To more directly assess the value repositories generate for competitors, we measure the stock returns of a firm's closest peers around repository releases. Using this approach, we

find that repositories generate only modest peer value. In fact, firms capture more value from their repositories than their 30 closest peers combined. Thus, open-source innovation significantly predicts firm growth but generates little, if any, positive externalities for competitors, despite being made freely available. These findings extend the Schumpeterian model of growth and creative destruction to the open-source setting.

Altogether, the evidence indicates that competition is an important factor shaping how firms engage in open-source innovation. Firms facing less competition are more likely to share innovation via open source and derive more private value from that innovation. Managers and analysts often discuss open-source activities in the context of competition. Open-source innovation generates limited peer value, and licenses that restrict commercial use generate even more private value and even less peer value. These results suggest that firms are more likely to release innovation via open source when potential spillovers to competitors are less concerning. More broadly, open-source innovation by publicly traded firms does not appear to contribute to the ideal of open-source systems fostering industry-wide growth.

1.1 Related Literature

This paper makes several contributions to the literature on innovation. First, our paper contributes to the broad literature on measuring the economic value of innovation. Existing studies have explored innovation within traditional intellectual property systems, such as patents or trademarks, which grant exclusive rights to use and monetize innovative outputs (Pakes, 1985; Austin, 1993; Hall et al., 2005; Kogan et al., 2017; Chen et al., 2019; Desai et al., 2025; Ahmadi et al., 2024; Liu et al., 2024). In contrast, we explore how innovation contributes to firm value even when freely disclosed under open-source licenses. In this setting, as noted by Lerner and Tirole (2005a), value is indirectly generated, making it challenging to measure quantitatively. We address this challenge by leveraging financial markets to measure the value of intellectual property without legal excludability.

Most directly, our paper contributes to the literature on open-source innovation. We

construct an extensive dataset of open-source activity by public firms on GitHub, which allows us to document open-source activity at a granular level. It also allows us to develop a new stock-market-based measure of the value of open-source innovation. Previous research has estimated the value of specific open-source software, such as Apache and nginx, using the cost to replicate similar services with proprietary software (Greenstein and Nagle, 2014; Murciano-Goroff et al., 2021). Other research estimates the aggregate economic value of open-source software using a cost-to-produce approach based on code length and labor costs (Robbins et al., 2021; Blind et al., 2021). For example, Hoffmann et al. (2024) estimate the cost of replicating the most-used open-source software either once ($4.15 billion) or individually by all users ($8.8 trillion).[6] In contrast, we measure the private value of open-source activity by public firms using stock-market reactions, allowing us to quantify the dollar value of individual repositories and explore heterogeneity therein. This approach has the advantage of capturing indirect channels of value, which is essential for measuring the private value of open-source innovation, and provides an ex-ante measure of repository quality.

We also contribute to the literature on the sources of private value from open-source innovation, which we survey in more detail in Internet Appendix A. We group these sources into three broad mechanisms: value generated through adoption, value generated through community, and value affected by competition.[7] Our analysis provides evidence on a variety of channels within these mechanisms. Most notably, our results speak to the importance of competition in determining the value generated by open-source innovation.

[6] Outside of the realm of open source, Gómez-Cram and Lawrence (2025) investigate the value of software by analyzing the long-run stock returns of software firms.

[7] These mechanisms are drawn from a broad literature that considers the possible incentives for freely revealing innovation. These papers include Allen (1983), Lerner and Tirole (2002), Harhoff et al. (2003), Lakhani and von Hippel (2003), O'Mahony (2003), West (2003), Henkel (2006), Jeppesen and Frederiksen (2006), Lerner et al. (2006), von Krogh and von Hippel (2006a), Lakhani and Wolf (2007), Alexy et al. (2009), Henkel (2009), Dahlander and Gann (2010), Baldwin and Clark (2011), Casadesus-Masanell and Llanes (2011), Boudreau (2012), Henkel et al. (2014), Parker et al. (2017), Alexy et al. (2018), Nagle (2018), Nambisan et al. (2018), Teece (2018), van Angeren et al. (2022), and Lin and Maruping (2022). For reviews of the open-source literature, see von Krogh and von Hippel (2006b), Goldfarb and Tucker (2019), and Dahlander et al. (2021).

Our paper is most closely related to contemporaneous work by Mkrtchyan et al. (2025) and Coleman et al. (2025). The former studies how firms incorporate *external* innovation through open activities (e.g., hackathons), and the latter examines whether open-source engagement improves firm performance, particularly through future earnings, by complementing existing innovation via managerial learning and productivity channels. In contrast, we introduce a market-based, repository-level measure of the private value of open-source innovation, which enables exploration of a wide range of research questions that make use of the granular data available through GitHub.[8] Our approach facilitates analysis of project heterogeneity, licensing choices, and spillovers to peer firms, offering a broad perspective on both private and public value.

Finally, we contribute to the literature on innovation and firm growth (e.g., Aghion and Howitt (1992), Klette and Kortum (2004), Lentz and Mortensen (2008), Acemoglu et al. (2018), and Garcia-Macia et al. (2019)) by showing that the value derived from open-source innovation provides significant insight into firm growth beyond what is captured by patent value (Kogan et al., 2017). Previous research has shown that companies can enhance their software development capabilities, firm productivity, and access to venture capital by being active on open-source platforms (Nagle, 2018, 2019; Conti et al., 2021). Our study adds to this literature by testing the impact of open-source innovation on both a firm's and its competitors' long-term growth.

2 Institutional Background

In this section, we discuss the process of developing open-source innovation on the GitHub platform and provide institutional details necessary for understanding the data used in our analysis.

[8] Examples include: air quality and worker productivity (Holub and Thies, 2025), cybersecurity and cryptocurrency returns (Huang and Yang, 2025), generative AI and labor (Ye, 2025), labor reallocation (Gupta et al., 2025), diversity and productivity (Heath et al., 2025), and remote work (McDermott and Hansen, 2025).

2.1 Initiating Open-Source Projects

To deploy their projects on GitHub, firms need to create organization accounts. While some firms create only one account, others create multiple accounts based on organizational divisions, purposes, or related products. Repositories (projects) can then be created within these accounts, and administrators decide whether the projects will be publicly visible or only visible privately to certain organization or project members with the necessary permissions. The creation and management of public repositories come with almost no costs, while support and some features for managing private repositories require GitHub Team or GitHub Enterprise subscriptions.

One important decision to make when creating a repository is choosing a license. Without a license, projects cannot be considered open source, even if the source code is publicly visible.[9] The choice of license can have different implications for commercial use. There are two primary categories of licenses based on their permissiveness: permissive licenses and restrictive licenses. Permissive licenses impose minimal restrictions on how the source code can be used. In contrast, restrictive licenses limit how the code can be redistributed or combined with proprietary software. For example, copyleft licenses, which are one type of restrictive license, require that (part of) derivative projects using the licensed code must also be open source. Therefore, firms that intend to find a balance between sharing their work with the community and protecting their proprietary interests may find copyleft licenses more attractive, as their competitors may be hesitant to open source their proprietary developments built upon copyleft-licensed projects. In addition, some firms also opt for customized licenses with clauses that effectively limit commercial use. These custom licenses may appear open source but include restrictions that make the projects more "source available" rather than truly open source.[10] Firms may also adopt a dual licensing strategy, allowing users to pay a fee to remove restrictions on commercial use. Internet Appendix B compares different types

[9] https://choosealicense.com/no-permission/

[10] https://opensource.org/osd/

of licenses based on permissions and conditions.

2.2 Project Development and Community Interaction

GitHub operates on the Git system, a collaborative and distributed platform for software development. In this context, several processes and community interactions play a central role in fostering innovation and progress. This section provides a brief overview of these processes.

The development process begins with the creation of a repository, where developers work on the code locally on their own computers. Changes are saved using the "commit" command, which records updates to the local repository along with brief summaries describing the modifications. Each commit serves as a checkpoint, documenting what was done and why. When developers are ready, they "push" or upload these commits to the remote repository. If the developer chooses to make their repository public, they may either develop the code within the repository before making it public or add code to the repository immediately after making it public.[11] In virtually all cases, however, the developer publicizes the repository with functioning code, as otherwise there is little incentive from community members to adopt the repository and contribute to its further development. This means that virtually all public repositories represent a functioning product when made public and exhibit a change from closed to open source.

Users interested in staying updated on a repository's progress can "star" a repository, essentially bookmarking it for future reference. Those who have questions or suggestions can also "open issues." Both the development team and fellow community members actively participate in addressing these issues.

Furthermore, users can engage in the development process by "forking" the repository, which allows them to create a personal copy and work on the codebase independently. If

[11] In our sample, 60% of repositories were created at least one week before their public release, with a mean of 117 days (median 42 days) prior to release. Since firms are unlikely to wait this long to release a repository if it were empty, this suggests that projects are largely developed before being made public.

the changes made in this personal fork are deemed valuable and applicable to the original project, users can initiate "pull requests." These pull requests serve as formal requests to integrate the changes back into the original repository. The changes proposed in pull requests undergo review and, if approved, are merged into the main codebase, thereby contributing to the open-source project's ongoing development.

3 Open-Source Activity

3.1 Data

To construct our dataset of GitHub activities by U.S. public firms, we begin by linking GitHub organization accounts with firms. We first collect websites of organization accounts via the GHTorrent project and the GitHub API. We then compare these domains with the web URLs of U.S. public firms and their subsidiaries from Compustat or Orbis. To ensure the accuracy of our matches, we screen out accounts whose domains are indicative of hosting or social media services, such as "github.com" and "facebook.com." We then conduct a rigorous manual search to complement our domain-based matching. Specifically, we query the firm names together with the term "open source" via Google to locate official web pages that list their open-source projects, and search the firm names on GitHub to identify associated organization accounts. Following this, we compile a comprehensive list of public repositories tied to the identified organization accounts through the GHArchive database, which records and archives timestamped public activity of GitHub repositories. In total, we match 1,281 firms with 3,314 organization accounts and 168,080 public repositories up to the year 2023.

Upon establishing a link between U.S. public firms and their respective GitHub organization accounts and public repositories, we utilize the GHArchive to gather additional information on the public footprints of these repositories. Most importantly, we determine the dates when the repositories were made public by identifying timestamps associated with the earliest activity, specifically those labeled as "PublicEvent." Pinpointing the exact dates

is crucial for our valuation process, which ultimately depends on the stock market reaction. We also create a firm-month panel that includes measures of aggregated activities observable to the public, such as the cumulative counts of repositories and the number of opened issues. Our panel spans the years 2015 to 2023, representing a relatively comprehensive picture of organizational engagement within the open-source community.

Additionally, we employ the GitHub API to collect static characteristics of 142,409 repositories extant as of February 2024. This includes an array of attributes from descriptive repository metadata, such as creation dates, licenses,[12] and programming languages, to quantitative measures of community engagement, including the number of stars and forks.

Finally, we use a large language model to classify or evaluate repositories based on topics, complementarity, and novelty. We use OpenAI's API to interact with the GPT-4o model, providing information including the repository name, description, main programming language, self-reported topics, website, and the name of the repository owner. We then prompt the model to conduct evaluation tasks. For topics, we use the model to assign a relatedness score (0 to 1) to 17 pre-defined topic categories, constructed from the GitRanking taxonomy (Sas et al., 2023). We define the complementarity score (0 to 1) as the extent to which a repository complements the firm's commercial products (instead of being a standalone product). We also use the model to evaluate the novelty of a repository (0 to 1), which measures how novel or groundbreaking it is compared to existing solutions, focusing on whether it introduces new ideas, techniques, or approaches. We take various steps to ensure consistency across repositories, including clear definitions of evaluation tasks, scoring reference systems, and consistency checks by conducting multiple rounds of scoring on a small subsample to verify stable outputs for each repository. Internet Appendix C provides details of our approach, model parameters, and prompts used.

[12] For customized licenses, we use large language models to determine whether they are permissive or restrict commercial use.

3.2 Summary Statistics

Before estimating open-source value, we first provide an overview of open-source activity. Figure 1 plots trends in open-source engagement during our sample period. The dashed yellow line shows the cumulative number of repositories created by public firms, totaling 122,971 by the end of the sample.[13] The blue line shows the percentage of public firms with at least one GitHub repository ("open-source firms"), rising from 4.8% in January 2015 to 18.2% in December 2023. The red line shows open-source firms' share of total market capitalization, reaching 67.5% by the end of the sample. The green line shows open-source firms' share of total research and development (R&D) expenditure, reaching 80.2% by the end of the sample. Thus, despite being only one-fifth of public firms, open-source firms represent two-thirds of the stock market and over four-fifths of innovation investment. Firms engaged in open-source activities are therefore an important part of the U.S. economy.

Next, we examine the distribution of open-source firms across industries. Figure 2 presents pie charts at the firm and repository levels. We use the Fama-French five industry classification scheme (Fama and French, 1997) and further separate "Computer Software" and "Finance" industries as defined by the Fama-French 49 industry classification scheme. Interestingly, only 32.2% of open-source firms come from the "Computer Software" industry. However, more than two-thirds of repositories are owned by firms in this industry. Nonetheless, other industries are also well represented in our sample, reflecting the growing importance of software across all parts of the economy.

We provide further summary statistics of open-source activities in Table 1. Panel A reports the distribution of repositories for all firms, as well as by industry, as of December 2023. Open-source activity is most common among "Computer Software" firms (66.2%), and least common in the "Healthcare, Medical Equipment, and Drugs" industry (7.5%).

[13] Cumulative counts in our firm-month panel are slightly smaller than in the original sample because we exclude (1) repositories never appearing in major open-source event records and (2) delisted firms starting from the month of delisting.

The latter may reflect the importance of excludability for innovation in this industry (Heller and Eisenberg, 1998).

Panel B compares firm and product market characteristics for firms with and without open-source activity over our sample period. Firm characteristics include the number of employees, market-to-book ratio, return-on-assets, investment, sales growth, tangibility, and R&D expenditure scaled by total assets, all of which are calculated using data from Compustat. We also calculate market capitalization and annual returns using data from the Center for Research in Security Prices (CRSP) and obtain patent data from Kogan et al. (2017).

We also examine the product market characteristics of open-source firms because technological spillovers to competitors represent the largest potential negative externality for open-source firms. These characteristics include market power, scope, product market centrality, product market similarity, and product market fluidity. Market power measures a firm's pricing power using a structural estimate of markups from Pellegrino (2025). Scope measures the number of industries in which the firm operates, based on product descriptions in SEC filings (Hoberg and Phillips, 2025). Product market centrality is calculated as eigenvector centrality in the product-market network, which is constructed using similarity scores from Hoberg and Phillips (2016). Central firms face more competition, but also benefit more from spillover effects in the network (i.e., network effects), which can be vital for driving adoption of open-source projects. Product market similarity measures how similar a firm's products are to its peers' (Hoberg and Phillips, 2016). Finally, product market fluidity measures how intensively a firm's product market is changing (Hoberg et al., 2014). A description of each variable and its data source is provided in Table A1.

We find that open-source firms are considerably larger on average, based on market capitalization, employees, and number of patents. They also tend to have higher valuations, based on market-to-book ratio, which could reflect investors' assessment of growth opportunities resulting from innovation. Intuitively, open-source firms have less tangible assets and larger R&D expenditures. Finally, open-source firms face less competition: they charge

higher markups, have lower product market centrality, are less similar to their product market rivals, and operate in less fluid product markets.

While these summary statistics paint a preliminary picture of open-source firms, they may be a function of other demographic differences between open-source and non-open-source firms, such as the concentration of open-source activity in the "Computer Software" industry. We therefore assess the correlation between characteristics and open-source activity in a regression setting in Internet Appendix D. Consistent with the summary statistics, open-source firms tend to be larger and more innovative, but they also have lower stock returns compared to their industry peers. We also investigate how the intensity of open-source activity, measured by the number of commits, relates to firm characteristics, providing an alternative to the binary comparison between open-source and non-open-source firms, and find similar associations. Importantly, these differences are largely absorbed when including firm fixed effects, suggesting that firm fixed effects can account for much of the selection bias in which firms choose to engage in open-source activities.

4 Open-Source Value

Having characterized open-source firms and assessed the determinants of open-source activity, we next estimate the economic value of open-source innovation by public firms. Specifically, we leverage financial markets to estimate the dollar value of GitHub repositories based on stock returns around the date the repository was made public. This estimate reflects the private value, which we define as the value captured by the firm, in contrast to the public value, which we define as the value to others, and the total value, which is the sum of the two. The estimate also reflects the value of the repository as a whole (i.e., the innovation value plus the value of being open source). It is also a forward-looking estimate as of the public release date, and as such does not capture changes in value that may occur as the project is further developed. However, as a forward-looking measure, it provides an ex-ante

measure of repository quality that is not a function of future information.

In measuring release returns, we assume that investors are aware of repository releases (or at least, awareness correlates with value). There are several channels through which investors may learn about repository releases. Many firms, such as Amazon, Microsoft, Apple, Salesforce, and Meta, maintain official websites where they announce new repositories. Prominent "tech" news websites, such as Wired, TechCrunch, and The Verge, post articles about particularly interesting, and potentially valuable, repositories. Useful repositories are also often shared on social media websites like Reddit, X/Twitter, and HackerNews.

However, even if investors are aware of repository releases, it is unclear whether the releases are viewed as relevant for firm value. Ultimately, this is a question of whether our estimates contain value-relevant information. To this end, we perform a validation exercise using future popularity in Section 4.3.1. To provide further qualitative evidence, Section 4.4 presents an analysis of earnings calls in which managers and analysts explicitly discuss open-source activities. We identify 988 such calls from 2015 through 2021, suggesting analysts view certain open-source activities as value relevant.[14]

We also assess the relevance of repository releases to investors by examining trading volume around releases. Specifically, we examine abnormal daily share turnover from three days before to three days after the release. Abnormal share turnover is measured by regressing turnover, calculated as trading volume divided by shares outstanding,[15] on indicator variables for each day relative to the repository release. The regression includes calendar-day fixed effects and firm-year fixed effects, and clusters standard errors by year-quarter. We also split the sample into large and small firms based on median market capitalization 10 days before the release. The coefficients estimated from this regression are plotted in Fig-

[14] For example, in the Qualys, Inc. earnings call for Q4 2021, an analyst asked about a trial version of a security application that had been posted on GitHub (https://github.com/Qualys/log4jscanwin): "You mentioned the 330-day web application scanning trial around Log4Shell in December in the prepared remarks. Just curious around the reception of that free service that drive new customers." The CEO responded: "Our goal definitely is to show the capability of the platform and how quickly we can spin up a new service and how quickly the customers can sign up and start getting value out of a cloud-based solution."

[15] To mitigate concerns about outliers, we winsorize share turnover at the 5% and 95% levels.

ure 3. For large firms, we find that share turnover increases around repository releases and is statistically significant on the day of the release. These results provide suggestive evidence, at least for large firms, that investors trade in response to repository releases.[16]

In the following, we outline the procedure to estimate repository private value, validate these estimates using a realized measure of repository popularity and a placebo test, and investigate the determinants of open-source value.

4.1 Estimating Private Value

Our procedure for estimating the private value of repositories closely follows the procedure developed by Kogan et al. (2017) to estimate patent value and used by Desai et al. (2025) to estimate trademark value. We provide a detailed discussion of this procedure in Internet Appendix E and briefly outline the crucial points in this section.

The procedure involves observing stock returns in the three-day window following the repository release, $[t, t+2]$.[17] We cumulate market-adjusted returns over the three-day event window for repository i, which we label R_i. We assume that R_i is a function of both investor reaction to the repository release, v_i, and idiosyncratic noise. We construct the estimate of repository value as the product of the investor reaction to the repository release and the firm's market capitalization on the day prior to the release. If multiple repositories are announced on the same day, we assume the value is evenly distributed across those repositories. The value of repository i, ξ_i, is thus calculated as

$$\xi_i = \frac{1}{N_i} E[v_i|R_i]M_i, \tag{1}$$

[16] Large firms represent 89% of repositories in our sample. Thus, the evidence suggests that investors respond to repository releases for the vast majority of our sample.

[17] While the positive coefficients for $t-2$ and $t-1$ in Figure 3 indicate there may be some information leakage before the release, we elect to proceed with the window $[t, t+2]$ to ensure our results are comparable to other studies using a similar methodology (Kogan et al., 2017; Desai et al., 2025). In untabulated regressions, we find quantitatively similar results for subsequent tests when extending the window to $[t-2, t+2]$.

where N_i is the number of repositories announced on that day, $E[v_i|R_i]$ is the expected return attributable to the repository release conditional on observing the three-day cumulative market-adjusted return R_i, and M_i is the market capitalization of the firm on the day prior to the repository release. We also use CPI values to adjust ξ to 2023 dollars. Internet Appendix E discusses our estimation of the conditional expected return in Equation (1), which adopts the same distributional assumptions as Kogan et al. (2017). Importantly, these assumptions imply that repositories have strictly positive values. We discuss this assumption further in the next section.

4.1.1 Comparing Open-Source and Patent Settings

It is worthwhile to briefly compare the validity of the methodological assumptions in the open-source and patent settings. First, since 2000, the USPTO publishes patent applications 18 months after filing, so part of the patent value is already reflected in the stock price before the grant date.[18] Kogan et al. (2017) adjust for this by estimating the probability of a patent being granted. In the open-source setting, while information leakage is possible before repository releases, such information is not systematically shared. This eliminates the need for a similar adjustment but means releases may also convey additional information, such as changes in innovation policy. We investigate this further in Section 4.3.1.

Second, patent grants are publicized by the USPTO every Tuesday. While this increases the potential for overlap in patent grants, managers cannot choose when their patents are granted. In contrast, repositories do not follow a systematic release schedule, reducing the potential for overlap in releases but also giving managers more discretion in when to release the repository. In Internet Appendix D, we find that repository releases do not consistently overlap with patent grants, earnings announcements, or trends in stock prices. We do find some overlap between repository releases and product releases and software-

[18] Note that the innovation quality is known to investors before the grant date, and thus the stock-price reaction reflects the value of receiving a patent on that innovation. However, because patent protection is more valuable when the innovation is more valuable, one still cannot separate the innovation value and the value of patent protection.

developer conferences, but the affected repositories represent a small fraction of our total sample (2.7% and 11.8%, respectively). Moreover, these repositories tend to be less valuable, meaning they are relatively under-weighted in our analysis. And even for repositories that overlap with these four types of events, we find that market reactions to repository releases contain value-relevant information. Thus, we conclude that the potential strategic timing of repository releases does not significantly bias our sample of value estimates. We also investigate whether managers' strategic behavior dilutes the information contained in release returns more generally in Section 4.3.1.

Third, Kogan et al. (2017) assume that patents have strictly positive values. This assumption provides structure to the distribution, allowing the release return to be separated into signal and noise, and reflects the idea that firms only apply for patents with a positive net present value. In the open-source setting, projects may benefit competitors in a way that reduces private value, but we still assume firms make projects open source only if the net effect is positive.[19] Nonetheless, we also provide an alternative calculation of private value,

$$\xi_i^{alt} = \frac{R_i M_i}{N_i}, \tag{2}$$

with summary statistics in Section 4.2, a validation test in Internet Appendix D, and a comparison to peer value in Section 5.1. While this alternative measure differs in the level of repository private value, it conveys the same information about the relative value of repositories in the cross section.

4.2 Summary Statistics

We estimate Equation (1) for the 28,905 *original* repositories with available release dates and required stock return data from CRSP.[20] Panel A of Table 2 reports the mean, standard

[19] Note that this assumes that repository private value is positive, not that release returns are positive. We discuss this assumption further in Internet Appendix E.

[20] We focus on original (i.e., not forked) repositories because release dates for forked repositories are not clearly defined.

deviation, and various percentiles (1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th) of several variables. The mean (median) three-day cumulative market-adjusted release return (R_i) is 0.12% (0.03%)[21] and the mean (median) expected return attributable to the repository release ($E[v_i|R_i]$) is 0.46% (0.39%). The difference between means and medians indicates positive skewness in market reactions.

The mean repository value, ξ, is \$849,443 and the median value is \$542,416. This variable is also significantly skewed: the 99th percentile of repository value exceeds \$5,000,000 and the most valuable repositories exceed \$12,000,000 (untabulated). These values are reported in 2023 dollars. For comparison, the mean value for patents granted over a similar period (2015-2023) is \$53 million in 2023 dollars, as calculated using data from Kogan et al. (2017). Given patents' higher investment costs and legal excludability, we consider the relatively lower repository value to be plausible. Moreover, as previously discussed, the types of innovation contained in patents and repositories are likely fundamentally different, with the alternative to open sourcing an innovation more often being trade-secret classification rather than patenting.

We also report statistics on repository characteristics. First, we report the number of stars each repository has received as of February 2024, which we use to measure the realized popularity of a repository. Again, we observe significant skewness in the variable. Second, we report complementarity and novelty scores on a 0-1 scale, reflecting how much a repository complements the firm's commercial products and its novelty relative to existing solutions. Most repositories (54.1%) significantly complement the firm's commercial products, with a complementarity score of at least 0.5. However, there is still a significant portion of standalone repositories (23.6% with a score of 0). Third, we report repository size, measuring bytes of data (code, images, etc.), which is also highly skewed. Finally, we report the cumulative number of issues opened per repository as of December 31, 2023. Issues typically convey community suggestions or error reports, but popular repositories are also more likely

[21] A t-test shows that the mean three-day cumulative market-adjusted release return is statistically significantly different from zero, with a t-statistic of 5.91.

to have issues opened in general.

Panel B summarizes repository values by industries. Most repositories come from the "Computer Software" industry (57.2%), while repositories from the "Consumer Durables, NonDurables, Wholesale, Retail, and Some Services" industry are the most valuable on average ($1,081,008). The "Healthcare, Medical Equipment, and Drugs" industry has the fewest repositories (97) and lowest rate of permissive licenses (45.4%), again consistent with the importance of excludability for this industry.

Panel C reports the 10 firms with the most valuable repository portfolios as well as the total value of all repositories in our sample. Amazon.com Inc. and Microsoft Corp. have the most valuable repository portfolios, each exceeding $7.5 billion. The remaining listed firms are also well-known technology firms with a focus on innovation, such as Alphabet Inc., Adobe Inc., and International Business Machines Corp. In total, repositories in our sample generated nearly $25 billion of private value for public firms.[22]

Panel D reports the 10 programming languages that generate the most value, based on each repository's main programming language. Python is the most common language (20.3% of repositories), has the most total value ($6,896,704,060), and has the highest average and median value among the languages listed.[23,24]

We also examine value across repository topics. Section 3.1 describes our procedure to assign topic scores, between zero and one, to each repository. To estimate how much value a repository generates for each topic, we multiply each topic score by the repository value.[25] Table 3 reports the mean, median, and total value generated by repositories with non-zero

[22] In Section 5.1, we provide a back-of-the-envelope calculation of the total value (private value + peer value) of all repositories released on GitHub from 2015 through 2023, which is approximately $909 billion.

[23] This statement includes Jupyter Notebook as a Python language because it is a web-based interface often used to work interactively with Python code.

[24] Other languages with a higher average repository value and at least 10 repositories include Cuda ($2,327,971, 23 repositories), Bicep ($1,468,167, 79 repositories), Swift ($1,435,616, 359 repositories), and CMake ($1,326,104, 25 repositories).

[25] Note that topic scores do not necessarily sum to one for a given repository, so these statistics should not be interpreted as a decomposition of repository value.

topic scores for each topic, along with the mean topic score, number of repositories, and share of repositories with permissive licenses.

"Core AI and ML" and "AI Applications" repositories are the most valuable on average, partly due to high average topic scores. However, even after adjusting for this (e.g., dividing Mean ξ by Mean Topic Score), these are still the most valuable topics. "Software Engineering" and "Cloud Infrastructure and DevOps" repositories generate the most total value, driven by most repositories being at least partially associated with these topics. "Advanced Data Analysis" is also notable for having a high mean value relative to its mean topic score.

Finally, we investigate how the average repository value has evolved over time. Figure 4 plots the average ξ, in 2023 dollars, of repositories released each quarter from 2015 through 2023. Average values hovered around $400,000 from 2015 through 2017, rose to between $600,000 and $800,000 from 2018 through 2019 (coinciding with Microsoft's acquisition of GitHub, announced in June of 2018), and increased again in early 2020, likely due to expectations of increased digitalization resulting from the COVID-19 shutdown. Since 2020, average values have hovered around $1,000,000, and most recently peaked above $1,300,000. Thus, the average repository value reported for the whole sample understates the value of open-source innovation in recent years.

4.3 Determinants of Open-Source Value

We next turn to explore which repository, firm, and product-market characteristics most strongly correlate with open-source value. Many of these characteristics overlap with each other, so we also include them together in regressions to assess their marginal correlations with open-source value. Internet Appendix F reports univariate correlations among all pairs of variables. We view these results as descriptive in nature and intended to provide a more complete characterization of open-source value.

4.3.1 Repository Popularity

We begin by investigating repository popularity. If release returns contain value-relevant information, then repository value should predict future repository popularity. Moreover, a significant correlation between these two variables puts an upper bound on any distortive effects arising from the concerns listed in Section 4.1. This includes concerns regarding whether investors pay attention to repository releases, other information conveyed by the release, strategic managerial behavior, and distributional assumptions, all of which predict a weak relation between repository value and future popularity.

To measure repository popularity, we use the number of stars each repository has received as of February 2024. "Starring" a repository bookmarks it for the user, allowing them to stay updated on the repository and indicating interest. We regress the natural logarithm of repository value on the natural logarithm of one plus the number of stars the repository has received.[26] We control for the natural logarithms of market capitalization (at release), volatility (over the release year), employees, and patent-portfolio value (both as of the prior year). We also include various fixed effects, including year, industry (three-digit SIC), repository topic, industry-year, firm, and firm-year. Standard errors are double-clustered by firm and year and all independent variables are standardized.

Table 4 reports the results. Column (1) includes year fixed effects and reports a significant correlation between repository value and future popularity. Column (2) adds industry and repository-topic fixed effects and Column (3) replaces industry fixed effects with industry-year fixed effects. In both cases, we continue to find a significant correlation between repository value and future popularity. Economically, the estimate reported in Column (3) indicates that repositories that end up being one standard deviation more popular have an 8.6% higher valuation when released. Finally, Columns (4) and (5) add firm and firm-year fixed effects. Both results demonstrate a strong correlation between within-firm repository value and future popularity.

[26] We log-transform each variable to adjust for the skewness documented in the previous section.

Internet Appendix D reports similar regressions with alternative specifications: replacing stars with forks; excluding Amazon.com Inc., Microsoft Corp., and Alphabet Inc. from the regression; using only release-day returns to estimate ξ; and replacing ξ with ξ^{alt}. All regressions show a positive and statistically significant correlation between private value and future popularity.

We therefore conclude that repositories estimated to be more valuable when they are released are significantly more popular in the future. While we cannot exclude the possibility that some repository values will be less informative due to the concerns enumerated in Section 4.1.1, this result indicates that the estimation procedure captures significant value-relevant information on average.

4.3.2 Placebo Test

To further validate our estimates of repository value, we conduct a placebo test using randomized release dates. We first randomly assign each repository a placebo release date in the same calendar year as the true release date. We then estimate the placebo value using stock returns on this placebo release date. Finally, we regress placebo value on the true number of stars the repository subsequently receives, following the specification in Column (5) of Table 4 (i.e., including repository-topic and firm-year fixed effects). We repeat this process 500 times.

Panel A and B of Figure 5 plot the distributions of the resulting coefficients and t-statistics, respectively, from these placebo regressions. In both panels, we include a vertical dotted line representing the corresponding estimates from the true release dates (i.e., Column (5) of Table 4). The figures show that the true coefficient and t-statistic are clear outliers relative to the placebo estimates. While four of the 500 iterations produce t-statistics of similar magnitude to the true relation, none of the iterations produce a coefficient approaching the true relation. We therefore conclude that the estimates of repository value contain value-relevant information.

4.3.3 Repository and Product Market Characteristics

We next investigate correlations between repository value and other repository and product market characteristics.[27] The results are reported in Table 5. Each regression includes repository-topic and industry-year fixed effects and controls for repository popularity, stock market capitalization, stock volatility, employees, and total patent value. Within each panel, characteristics are sequentially introduced, with the final column including all characteristics from that category.

Panel A examines repository characteristics. We first consider license type. The license of each repository is classified as either permissive (no restrictions on use) or restrictive (some restrictions on use). We include an indicator variable for restrictive repositories such that permissive repositories represent the omitted category. Theory suggests that excludability increases the private value of innovation (Schumpeter, 1912), and Lerner and Tirole (2005b) discuss how restrictiveness increases open-source value. Consistent with these theories, restrictive repositories are approximately 7.5% more valuable on average (Column (8)).

We next examine repositories designated as templates, which can be easily duplicated without keeping the commit history. Consistent with these repositories being less likely to contain a unique innovation, we find that template repositories are approximately 14.2% less valuable on average (Column (8)).

We also examine how repository value correlates with repository complementarity and novelty. Complementary repositories may be more valuable if they drive demand for commercial products, but standalone (i.e., less complementary) repositories may represent a more substantial innovation. Consistent with the latter, complementarity is negatively associated with repository value (Column (8)). Importantly, this result is robust to controlling for novelty, since complementarity may (inversely) capture an aspect of novelty, which is positively related to repository value (Columns (8)). Thus, complementarity does not appear

[27] Internet Appendix D reports regressions of repository value on firm characteristics. While future popularity is statistically significant in all specifications, no other firm characteristic is significantly correlated with repository value.

to be a first-order driver of repository value.[28]

We next examine repository size. Larger repositories have more lines of code, all else equal, and may therefore be more valuable. However, we find a negative relation between repository size and value. This could reflect other data, such as images, that increase repository size without additional lines of code. Examining the 2,223 repositories where size is separately categorized into binary (e.g., images) and non-binary (e.g., lines of code) data, we find that both data types are negatively related to repository value (untabulated).[29] We therefore conclude that larger repositories are not necessarily more valuable.

Finally, we examine the number of repositories released by the firm prior to the repository release date, and the cumulative number of issues opened for the repository as of December 31, 2023. The number of repositories is negatively and significantly related to value, suggesting an inverse relation between quantity and quality. The number of issues opened is also negatively and significantly related to value, which, given that the number of stars controls for popularity, suggests that repositories with more future bugs are seen as less valuable by investors at release.

Panel B investigates product market characteristics. These variables are of particular interest because a firm's product-market rivals are most likely to benefit from the open-source nature of a firm's repositories.

We first examine market power using the estimate of markups developed by Pellegrino (2025). Market power reflects a firm's pricing power and is negatively related to competition

[28] We cannot rule out the possibility that complementarity incentivizes firms to make highly complementary repositories open source. Thus, complementarity may still increase open-source value for a given repository, but only for those with high complementarity, which tend to be less valuable in the cross section. It is also possible that a firm's portfolio of repositories is made up of many high-complementarity repositories and relatively few standalone repositories, such that complementarity provides significant value across the firm's whole portfolio. However, we find that repositories with a complementarity score of at least 0.5 represent only 27.9%, on average, of the total value of firms' portfolios of repositories, and only 7.7% for the median firm. It therefore does not appear that complementarity represents a significant portion of the value of firms' repository portfolios.

[29] However, the ratio of non-binary size to total size is positively and significantly related to value (untabulated). It therefore appears that non-binary data (e.g., lines of code) contribute more to the value of a repository than binary data (e.g., images).

in theory. Firms with more market power may face less risk of competitors benefiting from their repositories, which could either incentivize them to open source more-valuable innovation or allow them to capture more of the total repository value. Consistent with these two possibilities, firms with more market power produce more-valuable repositories.

We then consider the relation between product market centrality and repository value. Centrality could reflect competition (suggesting a negative relation) or network effects (suggesting a positive relation). Consistent with these opposing predictions, centrality is insignificant when included by itself (Column (2)). However, it becomes positive and significant when including other product market characteristics that also reflect competition (Column (6)). This supports the hypothesis that network effects drive open-source value.

Finally, we also consider the scope, product market similarity, and product market fluidity of firms. Scope is negatively related to value, suggesting that firms with a sharper product focus produce more-valuable repositories. Product market fluidity is also negatively related to value, suggesting that repositories create more private value when product markets are more stable.

In summary, the results for product market characteristics suggest that repository value is negatively related to competition. This may be due to selection in the types of projects that are shared on GitHub: firms facing less competition may be more willing to share innovation that has a higher total value. Alternatively, these firms may be able to capture a larger fraction of the total value created by the repository. In either case, we find that competition is an important driver of the private value of open-source innovation.

Finally, it is important to note that repository popularity is positively and significantly related to repository value across all regressions reported in Table 5. This result further supports the validation exercises from the previous sections.

In Internet Appendix D, we report regressions of repository value on firm characteristics. While future popularity is statistically significant in all specifications, no other firm characteristic is significantly correlated with repository value. The lack of statistical significance

across firm characteristics after including our standard controls, particularly in contrast to the results for repository and product market characteristics, suggests these controls capture much of the firm-level variation in repository value. This finding narrows the scope for potentially omitted variables that could confound our analysis of firm growth in Section 5.

4.4 Management Discussion of Open-Source Value

The results in the previous section indicate that investors view competitive concerns as being particularly relevant for generating private value from open-source activities. In contrast, complementarity with commercial products is not as highly valued. In this section, we compare these results with managers' publicly stated views on how private value is derived from open-source activities.

To assess managers' views, we analyze firms' quarterly earnings calls. We first obtain transcripts of earnings calls from Refinitiv for 2002 through 2021. We then identify discussions of open source by filtering transcripts for the keywords "open source," "open-source," "open core," "open-core," "GitHub," and "GitLab." This results in 1,864 transcripts from 2002-2021 and 988 transcripts from 2015-2021.

We then investigate the extent to which the identified discussions align with various channels identified in the literature as potential sources of private value. Internet Appendix A provides a complete description of each of the 17 channels. We classify these channels into three broad groups: adoption, competition, and community. Given the evidence presented in the previous section, we pay particular attention to the Complementary Products channel, in which open-source projects increase demand for complementary proprietary products, and competition channels, which generate private value by affecting competitive positioning.

We use OpenAI's API to prompt the GPT-4o model to assess which channels align with firms' discussions of open source in earnings calls. First, the model determines whether the discussion conveys information about the private value generated by open source. Next, it determines whether the discussion concerns the release of, adoption of, or competition

induced by an open-source project. Finally, it identifies which channels of private value are explicitly discussed. Further details, including the full prompt, are in Internet Appendix C.

Table 6 reports the percent of earnings calls that explicitly discuss each source of private value. To match our sample in previous sections, we focus on transcripts from 2015 through 2021 that discuss private value. Panel A reports percentages for our three groups of channels. Adoption channels are most commonly discussed (59.4% of transcripts), followed by community (28.9%), and competition (25.8%). Competition channels are especially prevalent when the discussion concerns competition induced by open-source projects (62.1%).

Panel B reports percentages for each of the 17 sub-channels. Across all transcripts, the most commonly discussed channel is Complementary Products (25.4%). Thus, while managers most often highlight private value coming from complementarity, our previous results suggest investors are more skeptical of the value generated through this channel. Among competition channels, managers particularly discuss the use of open source to differentiate a product (17.7%) and undermine a competitor's market position (10.2%). It therefore seems that, in the context of competition, managers discuss open-source activities as a means to gain a competitive advantage, rather than the costs associated with potential spillovers. In contrast, our previous results indicate that investors are more concerned about the latter. This disconnect may stem from managers' incentives to offer an optimistic perspective on the firms' activities during earnings calls. In any case, these results demonstrate that managers make significant competitive considerations when forming their open-source strategy.

In Internet Appendix D, we investigate time trends in discussions of open source and the prevalence of the five most discussed channels: Complementary Products, Ecosystem, Product Differentiation, Community Development, and Market Growth. Overall, the number of discussions per year has increased substantially over time. Market Growth was the most discussed channel in the 2000's, but has since declined in favor of the Complementary Products and Ecosystems. Combined with our results highlighting the importance of competition for open-source value, this raises questions about whether open-source activities

contribute to firm or industry growth and how the resulting value is distributed. We explore these questions in the next section.

5 Open-Source Innovation and Firm Growth

Technological innovations are recognized as significant drivers of long-term firm growth (e.g., Aghion and Howitt (1992)). However, the financial gains or cash flows from such innovations often depend on the degree of excludability, enabled by protections such as patents or trademarks. In contrast, open-source licenses typically grant public usage rights, limiting firms' ability to directly appropriate returns. As a result, the link between open-source innovation and firm growth is uncertain.

At the same time, innovations by other firms are often associated with a negative impact known as creative destruction (e.g., Kogan et al. (2017)). This effect is grounded in the excludability provided by traditional intellectual property protection. By contrast, the public usage rights granted by open-source licenses may diminish creative destruction effects resulting from other firms' open-source innovation. However, the "free" nature of open-source innovation can also make it difficult for competitors to compete with similar technologies, creating competitive barriers and adding complexity to the relationship between firm growth, open-source innovation, and creative destruction.

To address these questions, we investigate whether open-source innovation predicts growth for the innovating firm or its competitors. We first calculate the firm-level repository value, $\xi_{f,t}$, as the sum of the repository values ξ_i for all repositories released by firm f in year t. We then scale $\xi_{f,t}$ with the concurrent total value of assets of the firm. Similarly, we calculate the asset-scaled aggregate value of repositories posted by firm f's competitors as the sum of repository values at the industry level (defined at the three-digit SIC level) in year t, excluding the repositories posted by firm f during the same period. We denote the firm-level repository value scaled by the firm's total assets as $Repo\ Output_{f,t}$ and the competitors'

repository value scaled by their total assets as $Repo\ Output_{I\setminus f,t}$, respectively.

We consider the growth of several dependent variables (Y), including sales, profits, the number of employees, and the value and number of patents:

$$\ln Y_{f,t+k} - \ln Y_{f,t} = \beta_k Repo\ Output_{f,t} + \gamma_k Repo\ Output_{I\setminus f,t} + \psi_k X_{f,t} + \epsilon_{f,t+k}, \tag{3}$$

where the horizon k varies from one to three years. The vector X includes the natural logarithms of one lag of the dependent variable ($Y_{f,t}$), firm capital, employment, idiosyncratic volatility, and the asset-scaled patent value of both the firm and its competitors each year. We also include industry (at the three-digit SIC level) and year fixed effects and double cluster standard errors by firm and year. To enable the comparison between open-source innovation output and patents, we standardize all independent variables. In a similar vein, we restrict our sample to firms for which all dependent variables are available for $k = 1$ to anchor our analysis on a balanced panel, while for longer horizons, we allow for unbalanced panels conditional on having complete data for $k = 1$.

The results are outlined in Table 7. The first three columns present our estimates of β_k for $k = 1, 2$, and 3, capturing the impact of a firm's open-source innovations on its growth over the next three years, respectively. We observe a significant positive relation between a company's open-source innovation output and its future growth. Economically, over the next three years, a one standard deviation increase in firm-level repository output is associated with a 2.2% increase in sales, a 1.9% increase in profit, and a 2.0% increase in employment. These effects are substantial: our estimates indicate that the growth impact of open-source innovation is somewhat comparable to that of patents, where a one standard deviation increase is associated with a 4.9% increase in sales, a 6.1% increase in profits, and a 4.3% increase in employment.

We also observe complementarity between open-source innovation and patentable innovation. A one standard deviation increase in firm-level repository output is associated with

a 4.4% increase in the value of new patents granted and a 3% increase in the number of new patents granted over the following three years. Furthermore, the larger coefficients for the patent value relative to the number of patents indicate that the average value per patent increases over time.

It is important to stress that we cannot make a causal conclusion about the impact of open-source innovation on firm growth from these results. For example, we cannot completely control for the time-varying innovative nature of firms, which may explain the similarity in economic magnitude of the effects of open-source and patented innovation despite large differences in their economic values. However, we do control for patented innovation in our regressions, so potentially omitted variables must reflect innovative characteristics that do not correlate with patenting. In any case, the results demonstrate that firms generating more private value from their open-source innovation experience more growth.

Turning to competitors' innovation, our estimates of γ_k capture the extent of creative destruction driven by open-source innovation. We observe that competitors' open-source innovation negatively affects a firm's long-term growth. Over the next three years, a one standard deviation increase in competitors' repository output is associated with a 1.7% decrease in sales, a 2.1% decrease in profits, a 0.7% decrease in employment, a 2.4% decrease in the value of new patents granted, and a 1.4% decrease in the number of new patents granted. These coefficients are comparable to those for patents, though slightly weaker. For sales growth, for example, the coefficients for competitors' patented innovation are statistically significant across all horizons, with a coefficient of 2.9% for a three-year horizon.

Additionally, we separate the private value generated by repositories with permissive and restrictive licenses, respectively, to investigate whether the negative externalities are stronger for repositories with more excludability. The results are reported in Table 8. Consistent with this hypothesis, we find that the negative externalities for all output variables are concentrated in repositories with restrictive licenses. By contrast, we find no such effects for repositories with permissive licenses and even observe positive externalities on sales.

Overall, our findings suggest that even when innovation is made accessible to others, firms still derive private value, providing strong evidence for why public companies engage in open-source innovation. Furthermore, our results indicate that the private value derived from open-source innovation by competitors negatively predicts a firm's growth. Once again, potential omitted variable bias prevents us from making causal conclusions from these results. Nonetheless, they are difficult to reconcile with the view that open-source innovation by public firms generates significant public value. We investigate this further in the next section.

5.1 Peer Value

In this section, we directly investigate the public value generated by open-source repositories. The public value of a repository includes the value it generates for other firms and non-commercial entities. While we cannot measure the value generated for all entities, we can estimate the value generated for publicly traded U.S. peer firms, which we call peer value. We thus investigate peer value as a window into the value public firms' open-source innovation generates for others.

To measure peer value, we would ideally apply our methodology for estimating private value to peer firms. However, this faces two main challenges. First, the distributional assumptions used to separate release returns into signal and noise, most notably that repository values are non-negative, are unrealistic for peer firms. To avoid these assumptions, we rely solely on release returns. While this yields a noisier estimate of value, it still offers insight into the peer value generated by repositories. Second, broadly defined peer groups can lead to many repositories being released by peer firms on the same day, increasing overlap in event windows and complicating the attribution of release returns across repositories. To mitigate this potential issue, we restrict the focal firm's peer group to certain sizes when calculating peer value.

To estimate peer value, we first calculate the three-day market-adjusted release return (i.e., R_i in Equation 1) for each peer firm following the repository release. We identify

peer firms using the product market similarity scores from Hoberg and Phillips (2016), which measure the similarity between firms' 10-K filings. We restrict the peer group to, at most, the 10, 30, 50, or 100 closest peers. Each peer firm's release return is then equally distributed across all repositories released by the peer firm's peers on the same day.[30] Finally, we calculate the value-weighted peer-firm release return for each repository. We also use the three-day market-adjusted release return for focal firms (i.e., R in Table 2) as a measure of private value to compare the private and peer values of each repository.

Table 9 reports summary statistics of the private and peer values of repositories. In the table, focal firms are indexed by i and peer firms are indexed by j. The first row of the table reports the average private value; the subsequent rows report the average peer value, based on different maximum peer group sizes.

The left side of the table converts returns into dollar values (i.e., ξ^{alt}). For peer value, these figures represent the total value generated for all peers in the peer group, averaged across repositories. While this conversion exacerbates noise in the peer value estimates, it allows us to calculate the share of total value (private + peer) captured by the focal firm.[31] On average, repositories generate \$339,273 of private value and between \$239,137 and \$455,408 of peer value. Peer value increases with peer group size, though the increase in value from 50 to 100 peers is small, suggesting an asymptotic peer value around \$500,000. These results imply that focal firms, on average, generate more value from releasing open-source repositories than their 30 closest peers combined and capture more than 40% of the total value generated for public U.S. firms across the broadest definition of peer groups.

The right side of the table reports the average release returns of repositories in our sample. For focal firms, the average return is 0.091%. For peer firms, it ranges from 0.050% to 0.032%, consistent with repositories generating more value for the focal firm than a typical peer firm. Additionally, the average peer-firm release return is largest when restricting the

[30] In unreported tests, we alternatively distribute returns based on peer scores or repository stars, and find qualitatively similar results.

[31] Note that "total value" in this context omits the value to private firms, non-U.S. firms, and individuals.

peer group to, at most, the 10 closest peers, suggesting that repositories generate more value for more-similar peers.

The final two columns of the table split the sample into repositories with restrictive and permissive licenses, respectively. Restrictive repositories have larger returns for the focal firm, though both license types are statistically significant. In contrast, peer firms have smaller returns for restrictive repositories. Moreover, these returns are statistically significant only for permissive repositories. This result is consistent with restrictive licenses allowing firms to capture a larger share of the total value generated by their repositories.

Finally, combining our estimates of total value across the broadest definition of peer group with the relation between private value and future popularity from Section 4.3.1, we conduct a back-of-the-envelope calculation of the total value generated by all repositories released on GitHub from 2015 to 2023. We begin by summing the total value and total number of stars (as of February 2024) for repositories in our sample released each year to compute a yearly value-to-stars ratio. Using GHArchive, we then sum the number of stars received by all repositories with an initial active date in each month and adjust this count using a conversion rate based on repositories appearing in both GHArchive and the GitHub API.[32] Finally, we cumulate the converted number of stars within each year and multiply by the value-to-stars ratio. This yields an estimated total value of approximately $909 billion for all repositories released on GitHub from 2015 through 2023.[33]

6 Conclusion

Given the importance of legal excludability in generating private value from innovation, the growing involvement of public firms in open-source innovation is initially surprising. To

[32] We make this adjustment because GHArchive records star additions but not removals.

[33] Again, this estimate of total value is based on the value a repository generates for a public firm and its peers. Thus, the omission of value for private firms, non-U.S. firms, and individuals induces a negative bias. However, it is also possible that repositories released by entities other than public U.S. firms have a lower value-to-stars ratio, which would induce a positive bias. The net effect is thus ambiguous.

explore this puzzling phenomenon, we construct an extensive dataset of open-source activities by public firms on GitHub, the largest open-source development platform, and use financial markets to develop a measure of the private value of open-source innovation.

We find that open-source engagement is highly prevalent in the U.S. economy. Firms with open-source projects tend to be larger, more valuable, and more innovative. Consistent with competitive pressures discouraging openness, firms facing less competition are more likely to engage in open source. Estimating the private value of open-source projects based on stock-market reactions, the average project in our sample is valued at $849,443 (in 2023 dollars), and the total value generated by all projects on GitHub from 2015 to 2023 is approximately $909 billion. Contrary to managers' discussions, projects that complement commercial products are seen as relatively less valuable by investors. Licenses that restrict commercial use and firms facing less competition generate more private value, further emphasizing the importance of competitive considerations. Valuable open-source innovation positively predicts future firm growth but negatively predicts competitors' growth. Directly measuring peer value reveals that firms capture more value than their 30 closest peers combined.

In summary, these results provide new evidence on the private value of innovation without legal excludability. The evidence shows that competition is a central concern for firms engaged in open-source innovation and suggests that firms may only release innovation via open source when they expect it to have limited value for competitors. More broadly, open-source innovation by publicly traded firms does not appear to contribute to the ideal of open source growing the market for all participants. Our estimates of open-source value open avenues for future research on the benefits firms derive from open-source innovation, its impact on competitors and the industry as a whole, and the valuation of intangible assets.

References

Acemoglu, D., U. Akcigit, H. Alp, N. Bloom, and W. Kerr (2018). Innovation, Reallocation, and Growth. *American Economic Review 108*(11), 3450–3491.

Acikalin, U., T. Caskurlu, G. Hoberg, and G. M. Phillips (2025). Intellectual Property Protection Lost and Competition: An Examination Using Large Language Models. Working paper.

Aghion, P. and P. Howitt (1992). A Model of Growth Through Creative Destruction. *Econometrica 60*(2), 323–351.

Ahmadi, A., A. Kecskés, R. Michaely, and P.-A. Nguyen (2024). Producing AI Innovation and Its Value Implications. Working paper, York University, University of Hong Kong.

Alexy, O., P. Criscuolo, and A. Salter (2009). Open for innovation: The role of openness in explaining innovation performance among uk manufacturers. *Strategic Management Journal 30*(12), 1318–1336.

Alexy, O., J. West, H. Klapper, and M. Reitzig (2018). Surrendering control to gain advantage: Reconciling openness and the resource-based view of the firm. *Strategic Management Journal 39*(6), 1704–1727.

Allen, R. C. (1983). Collective invention. *Journal of Economic Behavior & Organization 4*(1), 1–24.

Arrow, K. (1962). Economic Welfare and the Allocation of Resources for Invention. In *The Rate and Direction of Inventive Activity: Economic and Social Factors*, pp. 609–626. Princeton University Press.

Austin, D. H. (1993). An Event-Study Approach to Measuring Innovative Output: The Case of Biotechnology. *American Economic Review 83*(2), 253–258.

Baldwin, C. Y. and K. B. Clark (2011). The architecture of participation: Does code architecture mitigate free riding in the open source development model? *Management Science 57*(7), 1116–1133.

Blind, K., M. Böhm, P. Grzegorzewska, A. Katz, S. Muto, S. Pätsch, and T. Schubert (2021). The impact of Open Source Software and Hardware on technological independence, competitiveness and innovation in the EU economy. European Commission, Ed.

Bonaccorsi, A., S. Giannangeli, and C. Rossi (2006). Entry Strategies Under Competing Standards: Hybrid Business Models in the Open Source Software Industry. *Management Science 52*(7), 1085–1098.

Boudreau, K. J. (2012). Let a thousand flowers bloom? An early look at large numbers of software app developers and patterns of innovation. *Organization Science 23*(5), 1409–1427.

Casadesus-Masanell, R. and G. Llanes (2011). Mixed Source. *Management Science 57*(7), 1212–1230.

Chen, M. A., Q. Wu, and B. Yang (2019). How Valuable Is FinTech Innovation? *Review of Financial Studies 32*(5), 2062–2106.

Coleman, B., K. Fronk, and K. Valentine (2025). Corporate Adoption of an Open Innovation Strategy: Evidence from GitHub. Working paper.

Conti, A., C. Peukert, and M. Roche (2021). Beefing IT up for your Investor? Open Sourcing and Startup Funding: Evidence from GitHub. Accepted at *Organization Science*.

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

Dahlander, L. and D. M. Gann (2010). Open innovation in firms: A review of the literature. *Research Policy 39*(6), 699–709.

Dahlander, L., D. M. Gann, and M. W. Wallin (2021). How open is innovation? A retrospective and ideas forward. *Research Policy 50*(4), 104218.

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

Desai, P., E. Gavrilova, R. Silva, and M. Soares (2025). The Value of Trademarks. Working Paper, Nova School of Business and Economics.

Fama, E. F. and K. R. French (1997). Industry costs of equity. *Journal of Financial Economics 43*(2), 153–193.

Garcia-Macia, D., C.-T. Hsieh, and P. J. Klenow (2019). How Destructive Is Innovation? *Econometrica 87*(5), 1507–1541.

Goldfarb, A. and C. Tucker (2019). Digital Economics. *Journal of Economic Literature 57*(1), 3–43.

Gómez-Cram, R. and A. Lawrence (2025). The Value of Software. *American Economic Review forthcoming*.

Greenstein, S. and F. Nagle (2014). Digital dark matter and the economic contribution of Apache. *Research Policy 43*(4), 623–631.

Gupta, A., N. Nishesh, and E. Simintzi (2025). Big Data and Bigger Firms: A Labor Market Channel. Working paper.

Hall, B. H., A. Jaffe, and M. Trajtenberg (2005). Market Value and Patent Citations. *RAND Journal of Economics 36*(1), 16–38.

Hall, B. H. and M. MacGarvie (2010). The private value of software patents. *Research Policy 39*(7), 994–1009.

Harhoff, D., J. Henkel, and E. von Hippel (2003). Profiting from voluntary information spillovers: how users benefit by freely revealing their innovations. *Research Policy 32*(10), 1753–1769.

Heath, D., N. Seegert, and J. Yang (2025). Teamwork and the Homophily Trap: Evidence from Open Source Software. Working paper.

Heller, M. A. and R. S. Eisenberg (1998). Can Patents Deter Innovation? The Anticommons in Biomedical Research. *Science 280*(5364), 698–701.

Henkel, J. (2006). Selective revealing in open innovation processes: The case of embedded linux. *Research Policy 35*(7), 953–969.

Henkel, J. (2009). Champions of revealing—the role of open source developers in commercial firms. *Industrial and Corporate Change 18*(3), 435–471.

Henkel, J., S. Schöberl, and O. Alexy (2014). The emergence of openness: How and why firms adopt selective revealing in open innovation. *Research Policy 43*(5), 879–890.

Hoberg, G. and G. Phillips (2016). Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy 124*(5), 1423–1465.

Hoberg, G. and G. Phillips (2025). Scope, Scale and Concentration: The 21st Century Firm. *Journal of Finance 80*(1), 415–466.

Hoberg, G., G. Phillips, and N. Prabhala (2014). Product Market Threats, Payouts, and Financial Flexibility. *Journal of Finance 69*(1), 293–324.

Hoffmann, M., F. Nagle, and Y. Zhou (2024). The value of open source software. Working paper, Harvard University, University of Toronto.

Holub, F. and B. Thies (2025). Air Quality, High-Skilled Worker Productivity and Adaptation: Evidence from Github. Working paper.

Huang, D. and J. Yang (2025). Technology, Cybersecurity, and Cryptocurrency Returns. Working paper.

Jeppesen, L. B. and L. Frederiksen (2006). User communities and open innovation: The virtues of user innovation and its limitations. *Research Policy 35*(1), 1–16.

Klette, T. J. and S. Kortum (2004). Innovating Firms and Aggregate Innovation. *Journal of Political Economy 112*(5), 986–1018.

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

Lakhani, K. R. and E. von Hippel (2003). Hackers and the adoption of open source software. *Research Policy 32*(6), 923–943.

Lakhani, K. R. and R. G. Wolf (2007). Developers in the bazaar: The motivations and performance of open source software developers. *Management Science 52*(7), 984–999.

Lentz, R. and D. T. Mortensen (2008). An Empirical Model of Growth Through Product Innovation. *Econometrica 76*(6), 1317–1373.

Lerner, J., P. A. Pathak, and J. Tirole (2006). The Dynamics of Open-Source Contributors. *American Economic Review 96*(2), 114–118.

Lerner, J. and J. Tirole (2002). Some Simple Economics of Open Source. *Journal of Industrial Economics 50*(2), 197–234.

Lerner, J. and J. Tirole (2005a). The Economics of Technology Sharing: Open Source and Beyond. *Journal of Economic Perspectives 19*(2), 99–120.

Lerner, J. and J. Tirole (2005b). The Scope of Open Source Licensing. *Journal of Law, Economics, & Organization 21*(1), 20–56.

Lin, Y.-K. and L. M. Maruping (2022). Open Source Collaboration in Digital Entrepreneurship. *Organization Science 33*(1), 212–230.

Liu, L., E. Sojli, and W. W. Tham (2024). Firm Growth through Product Offerings. Working paper.

McDermott, G. R. and B. Hansen (2025). Labor Reallocation and Remote Work During Covid-19: Real-Time Evidence from Github. Working paper.

Mkrtchyan, A., J. Bai, R. Dai, and C. Wan (2025). Creativity without walls: The case of open innovation. Working paper.

Murciano-Goroff, R., R. Zhuo, and S. Greenstein (2021). Hidden software and veiled value creation: Illustrations from server software usage. *Research Policy 50*(9), 104333.

Nagle, F. (2018). Learning by Contributing: Gaining Competitive Advantage Through Contribution to Crowdsourced Public Goods. *Organization Science 29*(4), 569–587.

Nagle, F. (2019). Open Source Software and Firm Productivity. *Management Science 65*(3), 1191–1215.

Nambisan, S., D. Siegel, and M. Kenney (2018). On open innovation, platforms, and entrepreneurship. *Strategic Entrepreneurship Journal 12*(3), 354–368.

O'Mahony, S. (2003). Managing the boundary between firm and open source: From transparency to permeability. *Research Policy 32*(7), 1179–1198.

Pakes, A. (1985). On Patents, R&D, and the Stock Market Rate of Return. *Journal of Political Economy 93*(2), 390–409.

Parker, G., M. Van Alstyne, and X. Jiang (2017). Platform Ecosystems: How Developers Invert the Firm. *MIS Quarterly 41*(1), 255–266.

Pellegrino, B. (2025). Product Differentiation and Oligopoly: A Network Approach. *American Economic Review 115*(4), 1170–1225.

Robbins, C., G. Korkmaz, L. Guci, J. B. S. Calderón, and B. Kramer (2021). A First Look at Open-Source Software Investment in the United States and in Other Countries, 2009-2019. Paper prepared for the IARIW-ESCoE Conference.

Sas, C., A. Capiluppi, C. Di Sipio, J. Di Rocco, and D. Di Ruscio (2023). Gitranking: A ranking of github topics for software classification using active sampling. *Software: Practice and Experience 53*(10), 1982–2006.

Schumpeter, J. (1912). *The Theory of Economic Development*. Cambridge, MA: Harvard University Press.

Teece, D. J. (2018). Profiting from innovation in the digital economy: Enabling technologies, standards, and licensing models in the wireless world. *Research Policy 47*(8), 1367–1387.

van Angeren, J., G. Vroom, B. T. McCann, K. Podoynitsyna, and F. Langerak (2022). Optimal distinctiveness across revenue models: Performance effects of differentiation of paid and free products in a mobile app market. *Strategic Management Journal 43*(10), 2066–2100.

von Krogh, G. and E. von Hippel (2006a). Empirical research on open source software: The productivity effects of open source adoption. *Organization Science 17*(4), 497–514.

von Krogh, G. and E. von Hippel (2006b). The Promise of Research on Open Source Software. *Management Science 52*(17), 975–983.

West, J. (2003). How open is open enough? Melding proprietary and open source platform strategies. *Research Policy 32*(7), 1259–1285.

Ye, S. (2025). Task Efficiency and Signaling in the Age of GenAI: Effort Reallocation and Firm Value Effects. Working paper.

Figure 1
Trends in open-source engagement among U.S. public firms

This figure shows the time series of U.S. public firms' engagement in open-source activities from 2015 to 2023, based on the creation of public GitHub repositories. It plots the share of firms with public repositories as a percentage of total firms, total market capitalization, and total R&D expenditure (left y-axis). It also tracks the cumulative number of public repositories owned by these firms (right y-axis).

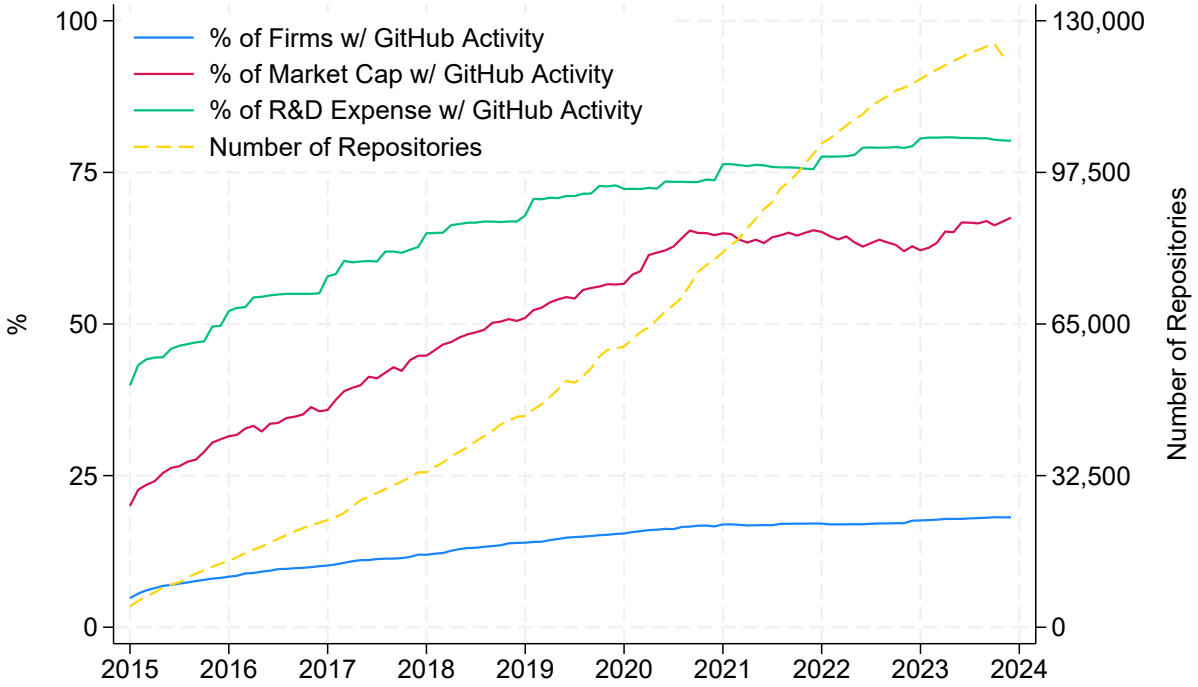


Figure 2
Industry distribution of open-source engagement

This figure presents the share of firms with GitHub activity and their associated repositories across industries from 2015 to 2023. The percentages are calculated by dividing the number of firms with GitHub activity in each industry by the total number of GitHub-active firms. The left pie chart shows the industry breakdown of firms with any GitHub activity, while the right chart depicts the distribution of their public repositories. Industry classifications follow the Fama-French 5 Industries, with Computer Software and Finance industries separated using the Fama-French 49 Industries.

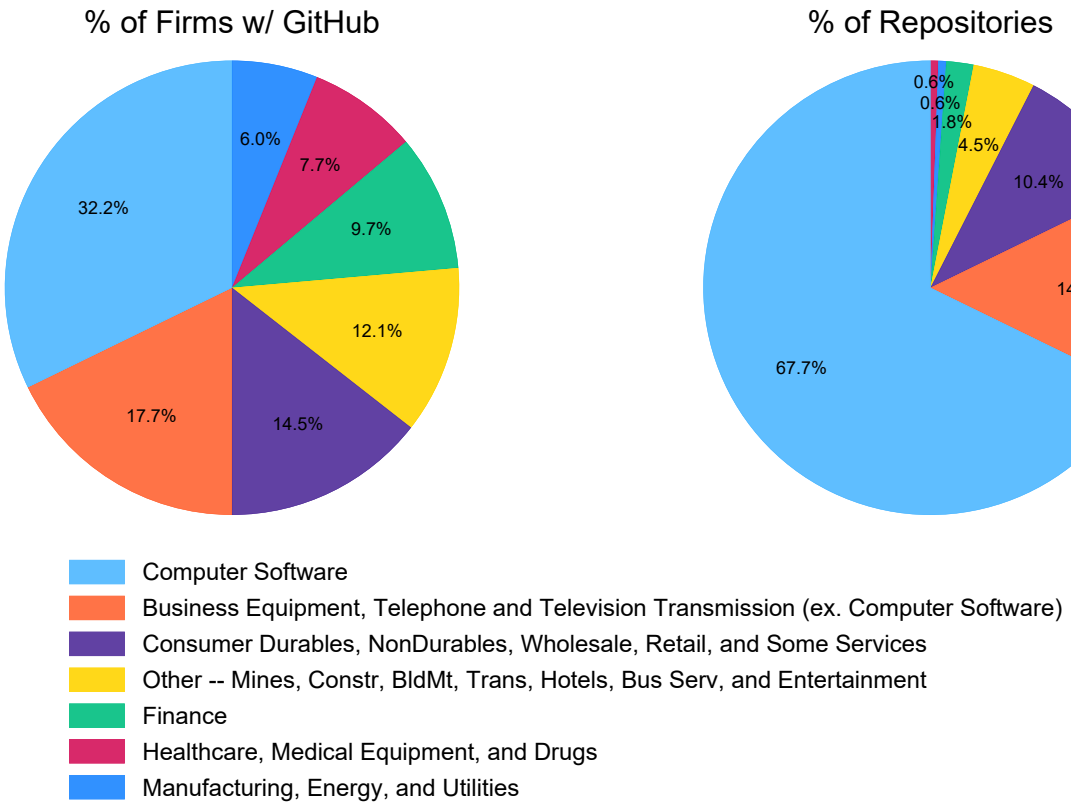
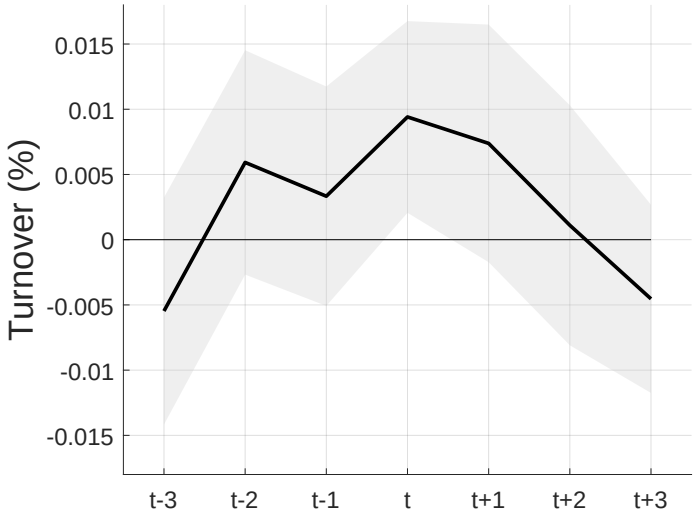


Figure 3
Share turnover around repository releases

This figure plots abnormal share turnover around repository releases. The values plotted represent coefficient estimates, along with 90% confidence intervals, from a regression of daily share turnover on indicator variables corresponding to three days before the release through three days after the release. The regression includes calendar day fixed effects and firm-year fixed effects and clusters standard errors by year-quarter. Panel A (B) reports results for the sample of firms with above-median (below-median) market capitalization ten days prior to the release.

Panel A: Large firms



Panel B: Small firms

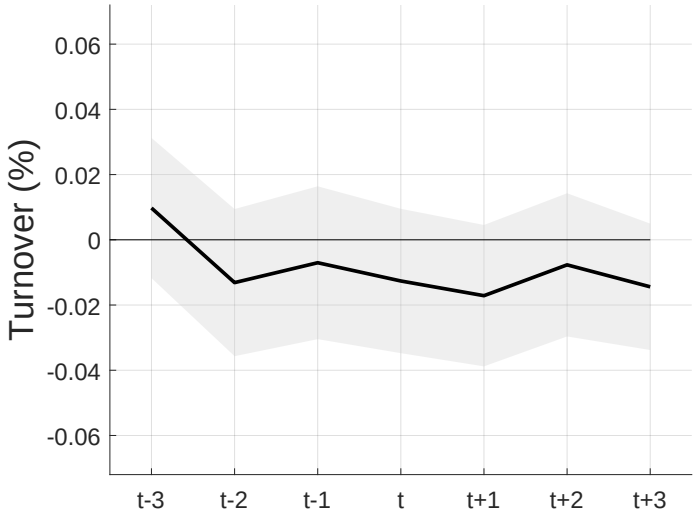


Figure 4
Average estimated repository value by quarter

This figure plots the average value, in 2023 dollars, of repositories released each quarter from 2015 to 2023.

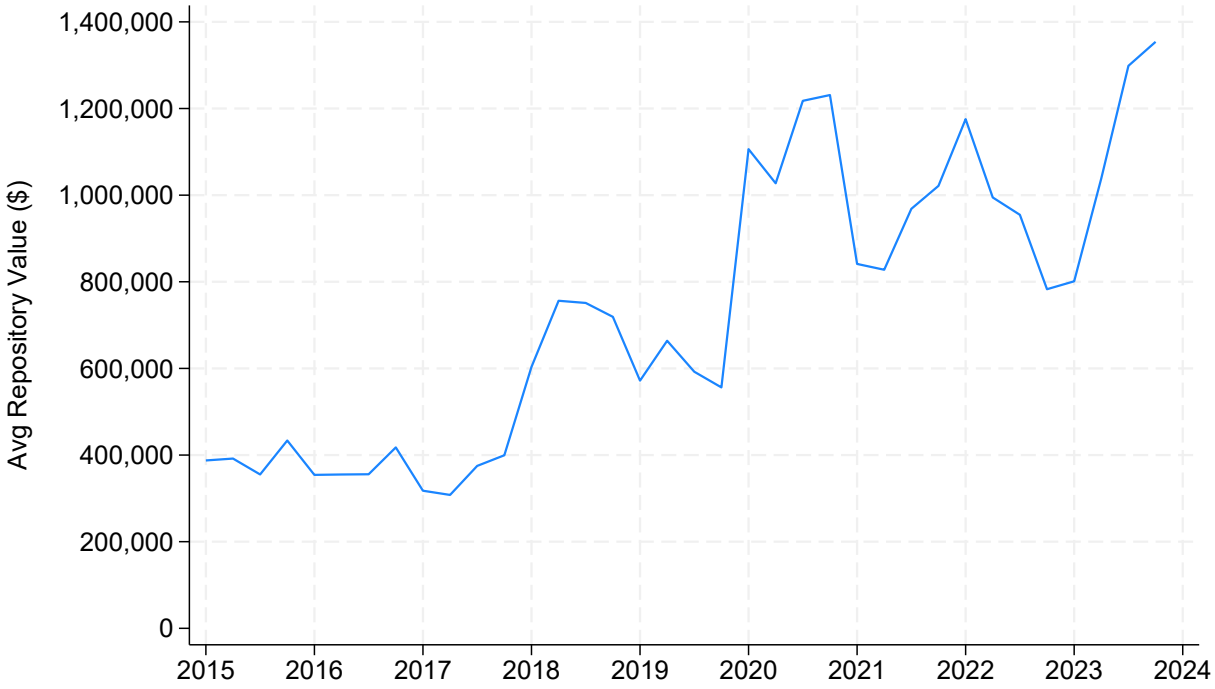


Figure 5
Placebo test results

This figure displays the distributions of coefficient estimates and t-statistics from 500 iterations of placebo tests. In each iteration, repositories are assigned a random placebo release date within the true release year. Repository placebo value is then determined by stock returns on that date. The upper figure shows the distribution of coefficient estimates from regressions of placebo values on the number of stars received, while the lower figure shows the distribution of corresponding t-statistics. Vertical dotted lines in both figures mark the actual coefficient estimate and t-statistics for comparison.

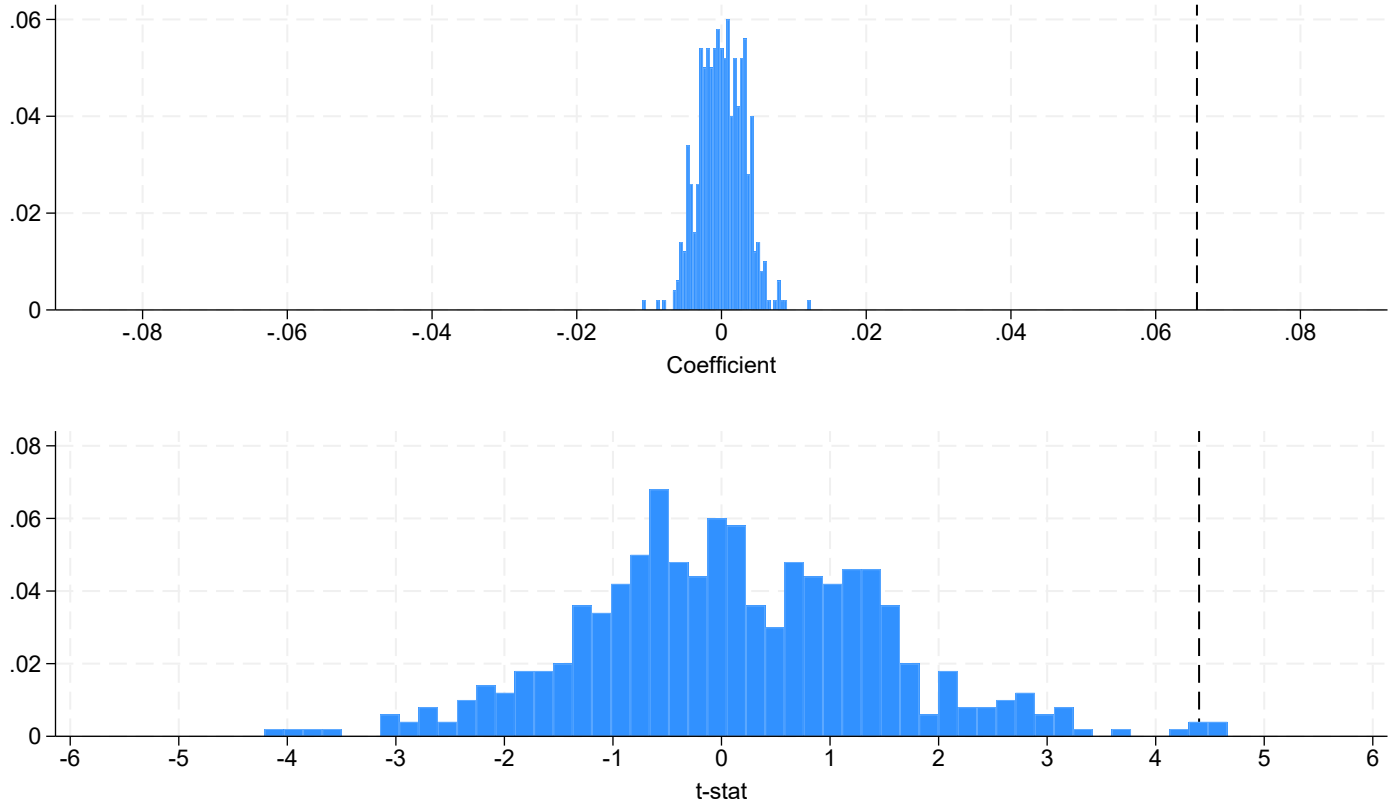


Table 1
Summary of GitHub Activity

This table presents summary statistics of GitHub activity. Panel A presents the percentage of firms engaged in GitHub activities and the distribution of repository ownership within each industry following the Fama-French 5 Industries, with Computer Software and Finance industries separated using the Fama-French 49 Industries. Panel B compares average financial characteristics, from 2015 to 2023, between firms that are active on GitHub and firms that are not. See Table A1 for variable definitions.

Panel A: GitHub Activity

		Number of Repositories									
	% GitHub	Mean	Std	p25	p50	p75	p90	p95	p99	Total	N Firms
Total	18.2%	30.9	430.1	0	0	0	17	62	425	122,971	3,982
Consumer Durables, NonDurables, Wholesale, Retail, and Some Services	20.8%	25.5	408.8	0	0	0	15	38	197	12,764	501
Manufacturing, Energy, and Utilities	7.8%	1.3	6.6	0	0	0	0	6	38	672	529
Business Equipment, Telephone and Television Transmission (ex. Computer Software)	34.7%	52.0	237.4	0	0	8	92	208	1253	17,535	337
Healthcare, Medical Equipment, and Drugs	7.5%	1.0	6.3	0	0	0	0	3	31	714	702
Other – Mines, Constr, BldMt, Trans, Hotels, Bus Serv, and Entertainment	22.2%	14.4	68.2	0	0	0	19	55	343	5,574	388
Computer Software	66.2%	246.4	1357.9	0	17	95	350	768	4551	82,288	334
Finance	10.9%	3.4	17.4	0	0	0	2	17	93	2,142	625

Panel B: Financial Statistics

	GitHub Firms		Non-GitHub Firms	
	Mean	Median	Mean	Median
Market Capitalization	32,221,598	3,754,909	1,417,841	71,486
Employees	33.0	4.9	8.1	1.2
Number of Patents	1,673	73	193	11
Market-to-Book	6.25	3.54	2.66	1.56
Return-on-Assets	-0.42%	2.48%	0.44%	3.51%
Investment	3.37%	2.09%	7.14%	4.45%
Annual Returns	14.44%	6.54%	15.21%	5.73%
Sales Growth	14.79%	9.21%	17.09%	9.34%
Tangibility	14.89%	9.04%	26.95%	20.42%
R&D Exp / Total Assets	8.03%	4.62%	3.81%	0.00%
Market Power	3.14	2.22	2.30	1.64
Scope	11	10	8	7
Product Market Centrality	0.0041	0.0021	0.0087	0.0038
Product Market Similarity	4.37	1.68	12.20	2.06
Product Market Fluidity	5.03	4.51	7.64	6.85

Table 2
Summary of Repository Value

This table reports summary statistics of repository private value for 28,905 repositories from 2015 to 2023. Panel A summarizes release returns, repository value, and other repository characteristics. R is the three-day cumulative market-adjusted release return. $E[v|R]$ is the conditional expected return attributable to the repository release. ξ is the estimated repository value reported in 2023 dollars. See Table A1 for definitions of the remaining variables. Panel B reports repository values by industry following the Fama-French 5 Industries, with Computer Software and Finance industries separated using the Fama-French 49 Industries. % Permissive is the percent of repositories with permissive licenses. Panel C lists the top 10 firms based on the total private value of their GitHub repositories. Panel D lists the 10 programming languages that generate the most private value.

Panel A: Summary Statistics

	Mean	Std	p1	p5	p10	p25	p50	p75	p90	p95	p99
R	0.12%	3.52%	-9.32%	-4.65%	-3.08%	-1.32%	0.03%	1.48%	3.45%	5.08%	10.34%
E[v|R]	0.46%	0.27%	0.18%	0.21%	0.24%	0.29%	0.39%	0.53%	0.79%	0.97%	1.39%
ξ	849,443	1,083,333	1,620	9,208	21,986	134,942	542,416	1,137,808	1,992,726	2,835,963	5,746,865
ξ^{alt}	220,774	9,221,580	-26,678,354	-10,131,999	-5,410,987	-1,306,678	2,407	1,622,631	6,260,536	11,630,696	27,924,340
Stars	216.5	2,250.8	0	0	0	3	10	45	215	578	3,832
Complementarity	0.42	0.28	0	0	0	0.1	0.5	0.7	0.8	0.8	0.8
Novelty	0.27	0.14	0	0.1	0.1	0.2	0.3	0.3	0.5	0.5	0.6
Repo Size	31,913.8	285,619.6	5	14	27	117	837	6,597	40,605	102,496	541,185
N Issues Opened	57.2	1,110.4	0	0	0	0	1	8	44	127	842

Panel B: ξ by Industry

	Total ξ	Mean ξ	Median ξ	N Repos	% Permissive
Computer Software	13,226,015,268	799,928	439,425	16,534	79.8%
Consumer Durables, NonDurables, Wholesale, Retail, and Some Services	7,842,714,887	1,081,008	885,356	7,255	92.3%
Business Equipment, Telephone and Television Transmission (ex. Computer Software)	3,078,350,930	904,068	208,129	3,405	58.5%
Other – Mines, Constr, BldMt, Trans, Hotels, Bus Serv, and Entertainment	149,470,183	390,262	242,453	383	76.8%
Finance	100,667,852	284,372	146,938	354	85.9%
Healthcare, Medical Equipment, and Drugs	51,090,658	526,708	267,219	97	45.4%
Manufacturing, Energy, and Utilities	47,465,899	280,863	248,078	169	76.3%

Panel C: Firms with Most Valuable GitHub Portfolios

	Total ξ	Mean ξ	Median ξ	N Repos	% Permissive
Amazon.com Inc.	7,747,034,059	1,135,429	916,747	6,823	94.1%
Microsoft Corp.	7,565,200,257	1,214,318	845,737	6,230	91.2%
Meta Platforms Inc.	2,260,058,454	1,931,674	1,722,887	1,170	54.4%
Alphabet Inc.	1,747,932,438	1,022,183	800,557	1,710	89.9%
NVIDIA Corp.	1,378,607,012	2,372,818	1,850,195	581	64.9%
Apple Inc.	1,049,619,695	4,447,541	3,667,135	236	65.3%
Salesforce Inc.	491,293,147	472,852	358,116	1,039	77.9%
Adobe Inc.	222,923,416	640,585	573,196	348	82.5%
International Business Machines Corp.	209,263,808	341,377	331,924	613	60.5%
Oracle Corp.	204,345,680	654,954	606,966	312	72.4%
...	...	...	...	...	...
Total	24,553,161,185	849,443	542,416	28,905	79.9%

Panel D: Most Valuable Programming Languages

	Total ξ	Mean ξ	Median ξ	N Repos	% Permissive
Python	6,896,704,060	1,173,308	820,856	5,878	81.9%
TypeScript	1,945,544,219	849,583	624,545	2,290	90.6%
JavaScript	1,675,214,332	544,785	257,651	3,075	82.0%
Jupyter Notebook	1,554,962,570	1,321,124	937,734	1,177	86.1%
C#	1,264,031,073	822,937	533,739	1,536	86.4%
Java	1,160,673,925	621,346	390,981	1,868	83.6%
C++	973,309,941	1,001,348	662,815	972	84.5%
Shell	785,121,777	803,605	550,226	977	78.9%
Go	759,058,408	525,664	220,817	1,444	87.5%
HCL	577,262,895	740,081	540,433	780	80.5%

Table 3
Repository Value by Topic

This table reports summary statistics of repository private value by repository topic. ξ is the repository private value reported in 2023 dollars. Topic Score, measured between zero and one, reflects how related a repository is to the topic. % Permissive is the percent of repositories with permissive licenses.

	Mean ξ	Median ξ	Total ξ	Mean Topic Score	N Repos	% Permissive
Core AI and ML	896,239	627,243	2,872,447,378	0.60	3,205	78.1%
AI Applications	837,759	586,353	2,878,539,523	0.58	3,436	78.1%
Digital Media	627,286	357,733	1,231,363,097	0.60	1,963	72.6%
Advanced Data Analytics	492,187	325,371	1,912,146,972	0.44	3,885	81.9%
Cloud Infrastructure and Development	488,057	357,248	5,334,954,346	0.56	10,931	88.4%
Security	432,695	260,176	1,455,153,834	0.53	3,363	84.7%
Education and Learning	409,770	231,314	640,470,744	0.46	1,563	77.5%
Configuration and Templates	378,219	257,167	872,173,835	0.45	2,306	85.6%
Development Tools	366,657	198,961	1,930,813,943	0.48	5,266	85.9%
OS and Platforms	359,838	174,177	681,174,079	0.47	1,893	72.0%
General Data Handling	322,041	208,009	2,426,899,944	0.38	7,536	83.9%
Software Engineering	314,823	199,277	5,381,893,972	0.40	17,095	83.4%
Community and Governance	283,236	126,083	93,184,578	0.32	329	71.1%
Documentation	282,593	153,613	1,194,236,743	0.39	4,226	75.5%
Back-End Web Development	282,076	155,907	509,711,615	0.46	1,807	77.2%
Front-End Web Development	274,237	126,913	596,466,503	0.47	2,175	77.7%

Table 4
Repository Value and Future Popularity

This table reports regression results to validate the measure of private value for repositories (i.e., the dependent variable, ξ). The variable of interest is $ln(Stars + 1)$. "Stars" is the number of stars as of February 2024. See Table A1 for the definition of variables. All independent variables are standardized. Standard errors double clustered by firm and year are reported in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)
ln(Stars + 1)	0.125***	0.104***	0.086***	0.074***	0.065***
	(0.032)	(0.023)	(0.018)	(0.014)	(0.015)
ln(Mkt Cap)	1.866***	1.857***	1.792***	1.892***	1.613***
	(0.156)	(0.123)	(0.110)	(0.084)	(0.151)
ln(Volatility)	0.377***	0.568***	0.592***	0.465***	
	(0.082)	(0.052)	(0.056)	(0.033)	
ln(Employees)	-0.334*	0.058	0.162	-0.180	
	(0.156)	(0.148)	(0.119)	(0.189)	
ln(Patent Value + 1)	0.097	0.036	0.058	0.115	
	(0.075)	(0.050)	(0.037)	(0.092)	
Observations	27,747	27,747	27,747	27,747	27,747
Adj. R^2	0.787	0.808	0.819	0.853	0.863
Year FE	✓	✓			
Industry FE		✓			
Repo Topic FE		✓	✓	✓	✓
Industry x Year FE			✓	✓	
Firm FE				✓	
Firm x Year FE					✓

Table 5
Determinants of Repository Value

This table reports which repository characteristics (Panel A) and product market characteristics (Panel B) are correlated with repository private value (ξ). Control variables include market capitalization, volatility, employees and patent value. See Table A1 for the definition of variables. All independent variables are standardized. Standard errors are double clustered by firm and year in Panel A and by industry and year in Panel B, and are reported in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

Panel A: Repository Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ln(Stars + 1)	0.089***	0.086***	0.085***	0.064***	0.094***	0.084***	0.146***	0.116***
	(0.017)	(0.018)	(0.016)	(0.018)	(0.016)	(0.019)	(0.025)	(0.027)
Restrictive License	0.087*							0.075*
	(0.039)							(0.037)
Template		-0.172*						-0.142
		(0.080)						(0.096)
Complementarity			-0.277***					-0.197**
			(0.073)					(0.074)
Novelty				0.363***				0.273***
				(0.074)				(0.076)
ln(Repo Size + 1)					-0.022**			-0.011
					(0.009)			(0.008)
ln(N Repos + 1)						-0.207*		-0.193*
						(0.105)		(0.100)
ln(N Issues Opened)							-0.081***	-0.055*
							(0.019)	(0.024)
Observations	28,061	28,061	28,061	28,061	28,061	28,061	28,061	28,061
Adj. R^2	0.825	0.825	0.826	0.825	0.825	0.827	0.826	0.829
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Repo Topic FE	✓	✓	✓	✓	✓	✓	✓	✓
Industry x Year FE	✓	✓	✓	✓	✓	✓	✓	✓

Panel B: Product Market Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
ln(Stars + 1)	0.073***	0.080***	0.072***	0.074***	0.075***	0.073***
	(0.011)	(0.011)	(0.011)	(0.010)	(0.011)	(0.011)
Market Power	0.086***					0.061**
	(0.019)					(0.021)
Product Market Centrality		0.036***				0.133***
		(0.006)				(0.019)
Scope			-0.110***			-0.073***
			(0.011)			(0.013)
Product Market Similarity				-0.070		-0.014
				(0.040)		(0.031)
Product Market Fluidity					-0.112***	-0.106**
					(0.015)	(0.038)
Observations	23,328	23,328	23,328	23,328	23,328	23,328
Adj. R^2	0.828	0.827	0.829	0.828	0.828	0.831
Controls	✓	✓	✓	✓	✓	✓
Repo Topic FE	✓	✓	✓	✓	✓	✓
Industry x Year FE	✓	✓	✓	✓	✓	✓

Table 6
Discussions of Open Source in Earnings Calls

This table reports how often firm managers mention private-value-generating channels in earnings calls related to open-source engagement from 2015 to 2021. Panel A groups channels identified in the literature into three broad categories: adoption, community, and competition. Panel B provides a breakdown of sub-channels. In both panels, the first column shows the share of transcripts related to each channel across the entire sample. The second through fourth columns report the share of transcripts based on different types of open-source engagement: releasing, adopting, or competing with an open-source project.

Panel A: Broad Channels

	All	Release	Adoption	Competition
Adoption Channels	59.4%	66.2%	64.1%	45.7%
Community Channels	28.9%	42.9%	29.5%	27.1%
Competition Channels	25.8%	20.2%	24.6%	62.1%
N Transcripts	716	317	468	140

Panel B: Sub-Channels

	All	Release	Adoption	Competition
Adoption Channels				
Complementary Products	25.4%	24.3%	27.8%	25.7%
Ecosystem	24.7%	26.2%	30.1%	18.6%
Market Growth	15.5%	24.0%	17.7%	18.6%
Network Effects	9.2%	14.5%	9.0%	2.9%
Technology Control	6.0%	8.8%	6.4%	2.1%
Community Channels				
Community Development	17.5%	31.9%	19.4%	15.7%
Transparency	4.1%	5.0%	4.9%	3.6%
Reputation	3.1%	4.7%	1.7%	2.1%
Integration Costs	2.2%	1.3%	3.0%	2.1%
Quality Certification	2.0%	1.6%	1.7%	5.7%
Identifying Talent	1.1%	1.9%	1.1%	0.0%
Employee Reputation	0.7%	0.6%	0.2%	0.0%
Signaling Innovation Direction	0.6%	0.3%	0.4%	0.0%
Competition Channels				
Product Differentiation	17.7%	14.8%	17.9%	36.4%
Undermine Competitors	10.2%	6.9%	8.3%	40.7%
Inherent Excludability	3.4%	2.8%	3.0%	7.1%
Prevent Entry	0.4%	0.3%	0.6%	0.7%

Table 7
Repository Values and Firm Output

This table reports the relation between the value of all repositories posted by a firm and its competitors in year t (ξ) and the firm's future outcomes over horizons from year $t+1$ to year $t+3$, as specified in Equation 3. Dependent variables include the growth of: sales, profits, number of employees, number of patents, and patent value. Control variables include one lag of the dependent variable, firm capital, employees, idiosyncratic volatility, and the patent value of both the firm and its competitors. See Table A1 for the definition of variables. All variables are winsorized at the 1% level using annual breakpoints. All independent variables are standardized. Standard errors double clustered by firm and year are in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

	Firm (Horizon)			Competitors (Horizon)		
	1	2	3	1	2	3
Panel A: Sales						
Repo Output	0.009***	0.018**	0.022**	−0.004	−0.009	−0.017*
	(0.003)	(0.006)	(0.008)	(0.007)	(0.006)	(0.007)
Panel B: Profits						
Repo Output	0.008**	0.016**	0.019**	−0.007	−0.011	−0.021*
	(0.003)	(0.005)	(0.007)	(0.008)	(0.007)	(0.009)
Panel C: Employees						
Repo Output	0.007**	0.015**	0.020*	−0.008	−0.006	−0.007**
	(0.002)	(0.004)	(0.009)	(0.006)	(0.007)	(0.002)
Panel D: Value of Patents						
Repo Output	0.019**	0.035***	0.044***	−0.014	−0.033	−0.024
	(0.007)	(0.009)	(0.010)	(0.008)	(0.020)	(0.023)
Panel E: Number of Patents						
Repo Output	0.014***	0.021***	0.030**	−0.007	−0.016**	−0.014*
	(0.003)	(0.004)	(0.009)	(0.005)	(0.006)	(0.007)
Controls	✓	✓	✓	✓	✓	✓
Industry FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓

Table 8
Repository Values and Firm Output: Permissive vs. Restrictive Licenses

This table reports the relation between repository private value (ξ) and a firm's future outcomes over horizons from year $t+1$ to year $t+3$, as formulated in Equation 3. Dependent variables include the growth of: sales, profits, tangible assets, the number of employees, the number of patents, and patent value. Control variables include one lag of the dependent variable, firm capital, employees, idiosyncratic volatility, and patent value. See Table A1 for the definition of variables. Standard errors double clustered by firm and year are in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

	Competitors (Horizon)					
	Permissive			Restrictive		
	1	2	3	1	2	3
Panel A: Sales						
Repo Output	0.007	0.011	0.017*	−0.013**	−0.029*	−0.049***
	(0.004)	(0.007)	(0.008)	(0.005)	(0.014)	(0.007)
Panel B: Profits						
Repo Output	0.002	0.009	0.009	−0.007	−0.021	−0.034***
	(0.005)	(0.009)	(0.006)	(0.006)	(0.011)	(0.007)
Panel C: Employees						
Repo Output	0.004	0.007	−0.015	−0.013***	−0.016**	−0.017**
	(0.005)	(0.007)	(0.004)	(0.003)	(0.006)	(0.007)
Panel D: Value of Patents						
Repo Output	0.006	−0.008	0.050	−0.013	−0.028	−0.090*
	(0.014)	(0.031)	(0.030)	(0.009)	(0.021)	(0.035)
Panel E: Number of Patents						
Repo Output	−0.006	−0.008	0.015	−0.002	−0.010	−0.036***
	(0.004)	(0.010)	(0.007)	(0.006)	(0.011)	(0.009)
Controls	✓	✓	✓	✓	✓	✓
Industry FE	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓

Table 9
Private Versus Peer Value of Repositories

This table reports summary statistics comparing the private and peer values of repositories in 2023 dollars. The first row of the table reports the average private value of repositories following the same procedure as is used to calculate peer value. Focal firms are indexed by i and peer firms are indexed by j. R_i is the three-day market-adjusted return following the repository release. N_i is the number of repositories across which the release return is distributed. Columns labeled "All" report statistics for all repositories, while columns labeled "Restrictive" and "Permissive" report statistics for repositories with restrictive and permissive licenses, respectively. % Total is calculated as Peer Value divided by the sum of Peer Value and Private Value. VW R_j/N_j is the market capitalization-weighted R_j/N_j. N Peers $\leq X$ indicates the sample of peer firms is restricted to at most X peers. Standard errors are reported in parentheses below each statistic. ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

	ξ_i^{alt}		R_i/N_i		
	All		All	Restrictive	Permissive
Private Value	\$339,273***		0.091%***	0.105%**	0.086%***
	(63,142)		(0.019%)	(0.053%)	(0.020%)
Peer Value	$\sum_j \xi_j^{alt}$	% Total	VW R_j/N_j		
N Peers $\leq$ 10	\$239,137***	41.3%	0.050%***	0.035%	0.054%***
	(55,238)		(0.010%)	(0.023%)	(0.011%)
N Peers $\leq$ 30	\$317,616***	48.4%	0.032%***	0.013%	0.038%***
	(64,225)		(0.007%)	(0.017%)	(0.008%)
N Peers $\leq$ 50	\$435,976***	56.2%	0.033%***	0.009%	0.041%***
	(65,982)		(0.006%)	(0.015%)	(0.007%)
N Peers $\leq$ 100	\$455,408***	57.3%	0.033%***	0.014%	0.039%***
	(66,867)		(0.005%)	(0.013%)	(0.006%)
N Repos	22,397		22,397	5,139	17,258

Appendices

Appendix A

Table A1
Variable Definitions

Variable	Definition
Complementarity	Score between zero and one that measures how much the repository complments the firm's commercial products (ChatGPT).
Employees	Number of employees (Compustat).
Firm Capital	Property, plant, and equipment - total gross (Compustat).
Investment	CAPX scaled by lagged total assets (Compustat).
Market Capitalization	Share price times the number of shares outstanding (CRSP).
Market Power	An estimate of markups assuming constant returns to scale developed by (Pellegrino, 2025).
Market-to-Book	Ratio of market capitalization to book equity, where book equity is calculated following Davis et al. (2000) (CRSP, Compustat).
N Issues Opened	Cumulative number of issues opened for a repository as of December 31, 2023 (GHArchive).
N Repos (t)	Cumulative number of repositories released by a firm prior to month t (GHArchive).
Novelty	Score between zero and one that measures how novel or groundbreaking a repository is compared to existing solutions, focusing on whether it introduces new ideas, techniques, or approaches (ChatGPT).
Number of Patents	Number of patents granted (Kogan et al., 2017).
Patent Value	An estimate of the economic value of patents using stock market returns around the patent grant date (Kogan et al., 2017)
Product Market Centrality	Eigenvector centrality calculated from a network created by product market similarity scores (Hoberg and Phillips, 2016).
Product Market Fluidity	A measure of how intensively the product market around a firm is changes (Hoberg et al., 2014).
Product Market Similarity	A measure of how similar a firm's products are to its peers', from Hoberg and Phillips (2016) (Hoberg-Phillips Data Library).
Profits	Sale minus COGS, deflated by the CPI (Compustat)
R&D Exp/Total Assets	Research and development expense scaled by lagged total assets (Compustat).
Repo Size	Byte size of a repository as of February 2024 (GitHub API).

Continued on the next page

Continued

Variable	Definition
Repo Output	An asset-scaled estimate of the economic value of repositories (in 2023 dollars).
Restrictive License	Indicator variable that equals one if the repository has a license that restricts commercial use (GitHub API).
Return (t)	Returns from date t (CRSP).
Return-on-Assets	Net income divided by lagged total assets (Compustat).
Sales	Annual Sales (Compustat).
Sales Growth	Annual percentage change in sales (Compustat).
Scope	Number of industries in which the firm operates, see Hoberg and Phillips (2025) (Hoberg-Phillips Data Library).
Stars	Number of stars of a repository as of February 2024 (GitHub API).
Tangibility	Property, plant, and equipment scaled by total assets (Compustat).
Template	Indicator variable that equals one if the repository is configured as a template, which allows copies to be created without retaining the commit history (GitHub API).
Topic Score	Score between zero and one that measures how much the repository relates to the given topic (ChatGPT).
Turnover	Trading volume divided by shares outstanding (CRSP).
(Idiosyncratic) Volatility	Standard deviation of daily returns over one month. Idiosyncratic volatility is similarly defined using returns net of market returns (CRSP).
ξ	An estimate of the private value of repositories (in 2023 dollars) using stock market returns around the repository release date.
ξ^{alt}	An alternative estimate of the private value of repositories (in 2023 dollars) using stock market returns around the repository release date that does not make assumptions about the distribution of repository values.

Internet Appendix A

Channels of Private Value

The literature on open-source innovation highlights many channels through which firms can capture private value from making their innovation publicly available. In total, we identify 17 channels present in the literature and organize these channels into three fundamental categories. The first category generates value through increased adoption. By removing barriers to adoption, firms can accelerate the diffusion of their innovations, creating value through various forms of network effects. The second category generates value through the open-source community. These channels primarily operate through reputation effects, signaling mechanisms, and human capital considerations. The third category generates value through effects on competition. Firms may engage in open source to strategically position themselves relative to competitors, either through differentiation, market positioning, or competitive dynamics. We outline and discuss the channels constituting each category below.

Adoption Channels

- *Complementary Products*: Making an innovation open-source can increase demand for related proprietary products or services offered by the firm. In this way, firms deliberately subsidize one component to increase demand for profitable complements. The open-source component acts as an entry point, while the firm captures private value through increased sales volumes for complementary products. (Lerner et al., 2006; Casadesus-Masanell and Llanes, 2011; Henkel et al., 2014; Alexy et al., 2018; Teece, 2018)

- *Ecosystem*: Open sourcing a technology can help establish it as a component of a broader technological ecosystem, thereby increasing switching costs for users. By mak-

ing the technology open source, users are more likely to adopt the technology and join the firm's ecosystem. This relies on complementarity between the open-source project and other components of the ecosystem. As adoption grows, users and developers may build additional applications and tools around the open-source component, further enhancing the influence of the ecosystem and reinforcing network effects. As users become increasingly locked into the ecosystem, firms can monetize complementary components within the ecosystem, charge for premium services or support, or leverage ecosystem lock-in to maintain customer relationships for other products. (Parker et al., 2017; Lin and Maruping, 2022)

- *Market Growth*: Sharing innovation via open-source platforms can expand the overall market for a certain product or service, allowing all producers, including the firm and its competitors, to benefit from increased market size. By making their technology freely available, firms can accelerate market development, stimulate complementary innovation by other parties, and reduce barriers to adoption. Firms can particularly benefit from this channel if they can maintain competitive advantages in the expanded market. These advantages include superior implementation capabilities, better customer relationships, more effective marketing, or superior complementary assets. (Parker et al., 2017)

- *Network Effects*: These represent the most direct manifestation of adoption-driven value creation. Both direct and indirect network effects can be present. Direct effects arise when additional users enhance the utility of existing users, such as through accelerated development and testing. Indirect effects can manifest through complementary products and services that are developed as the user base expands. Monetization of network effects often takes place through other channels, such as complementary products or technology control. (Parker et al., 2017; Nambisan et al., 2018; Lin and Maruping, 2022)

- *Technology Control*: Increased adoption via open sourcing increases the likelihood that the technology becomes the standard for a particular task or industry. This grants influence to the firm over future technological development, such that the firm can shape the development to better match its idiosyncratic needs. This might include architectural decisions that favor the firm's complementary products, development priorities that align with the firm's strategic roadmap, or dependencies that leverage the firm's existing capabilities. Technology standards also have substantial switching costs, as users are reluctant to switch to competing technologies that may not be as widely accepted or accommodated by complementary tools. (Harhoff et al., 2003; Bonaccorsi et al., 2006; Teece, 2018)

Community Channels

- *Community Development*: Firms can benefit from community-driven contributions that reduce the resources required for software development. These benefits include software feature enhancements, performance improvements, and bug fixes, which often reflect real-world user experience and needs. Because such contributions come from a diverse group of external developers, they enable faster iteration, broader testing, and access to specialized expertise that can also stimulate organizational learning. As a result, firms may lower development costs and achieve higher software quality than would be possible through internal efforts alone. (Lakhani and von Hippel, 2003; Lerner et al., 2006; Dahlander and Gann, 2010; Nagle, 2018)

- *Transparency*: Open-source development fosters transparency by making source code publicly accessible, allowing users, developers, and other stakeholders to examine how the software functions. This openness reduces information asymmetries, builds trust, and lowers adoption barriers by ensuring there are no hidden functionalities or unknown risks. For firms, such transparency can reduce perceived risk and enhance credibility with both internal and external developers, investors, and regulators. The value of

transparency can be particularly significant in sectors where reliability and accountability are paramount, such as cybersecurity. By reinforcing the firm's image as open and dependable, transparency ultimately contributes to private value through greater stakeholder engagement and broader diffusion of the firm's technologies. (O'Mahony, 2003; Dahlander and Gann, 2010)

- *Reputation*: A strong reputation for innovation, openness, and technical leadership can create value for firms by enhancing customer loyalty, attracting talent, building investor confidence, and increasing influence within industry ecosystems. Active participation in open-source projects serves as a credible signal of these attributes, allowing firms to showcase their technological competence and collaborative approach. Such visibility reinforces trust and signals long-term viability, which can translate into reputational capital that differentiates the firm in both product and labor markets. (Henkel, 2006; Lerner et al., 2006; von Krogh and von Hippel, 2006a)

- *Integration Costs*: New employees who have previously contributed to or used the firm's open-source projects possess relevant human capital that transfers directly to their roles within the firm. This reduces the learning curve associated with firm-specific technologies and development practices. Thus, the firm benefits through reduced time and resources required to integrate new employees into the firm's system. (Boudreau, 2012)

- *Quality Certification*: Active maintenance and external developer engagement serve as credible signals of software quality and reliability, thereby creating value for firms. For users, visible community indicators, such as stars, forks, frequent updates, and transparent issue tracking, help reduce uncertainty about the technology's performance and long-term viability. These signals function as an informal form of quality certification, increasing user confidence and making the technology more attractive for adoption and integration. (Baldwin and Clark, 2011)

- *Identifying Talent*: Open-source participation enhances a firm's ability to recruit high-quality technical talent by reducing search costs and improving the alignment between developer capabilities and the firm's technical and organizational needs. Because contributions on open-source platforms are publicly observable, firms can evaluate developers' coding ability, technical interests, and collaborative behavior in real-world settings, offering a more reliable screening process than traditional hiring methods. This visibility allows firms to identify candidates who already align with their technological stack, problem domain, and team dynamics. In addition, the firm's active presence in open-source communities increases its visibility and credibility among potential employees, further expanding the pool of skilled applicants. (Lerner and Tirole, 2002; Jeppesen and Frederiksen, 2006; Lerner et al., 2006; Lakhani and Wolf, 2007)

- *Employee Reputation*: Employees may care about their reputation in the developer community, and so may request open sourcing certain projects to enhance their reputation. Firms may agree to these requests for high-quality employees as a form of compensation. Supporting employee participation in open-source projects can also create value for firms by enhancing their own reputation as technically sophisticated and developer-friendly. When employees gain visibility and recognition in open-source communities through their contributions, their individual reputations often reflect positively on the firm. These visible signals of technical excellence can also boost employee motivation, enhance job satisfaction, and improve the retention of top talent. (Lerner and Tirole, 2002; Henkel, 2009)

- *Signaling Innovation Direction*: Open-source innovation can create value by shaping external perceptions of a firm's strategic direction, both for developers and investors. Such decisions serve as credible signals of the firm's innovation priorities and long-term technological vision. This signaling also reduces information asymmetry for investors, who can better assess whether the firm is making meaningful technological progress

and allocating resources in promising directions. (Henkel, 2006; Alexy et al., 2009)

Competition Channels

- *Product Differentiation*: Firms may make their product open source to differentiate it from competitors' products. Open source can be a valuable product attribute for certain customers, particularly in markets where transparency is valued. Firms can increase their market share by tapping into customer segments with preferences for open source. (West, 2003; van Angeren et al., 2022)

- *Undermine Competitors*: In certain competitive circumstances, firms may need to undercut competitors' prices to attract customers. Sharing the product on open-source platforms, essentially setting the price to zero, is one way to do this. By offering free alternatives to competitors' paid products, firms can pressure competitor pricing and market share, undermining the competitor's market position. This strategy is particularly effective when the firm can also generate value from the open-source product through other channels and when the competitor holds a dominant position in a stagnant market. (West, 2003; Henkel, 2006)

- *Inherent Excludability*: Despite being publicly available, open-source projects may be difficult for competitors to use effectively. This can occur when projects require substantial complementary assets, specialized knowledge, or specific infrastructure that competitors cannot easily replicate. In this way, while the open-source project does not have legal excludability, the inherent characteristics of the project create excludability. For projects with a high degree of inherent excludability, firms can benefit from posting the project on open-source platforms through other value channels while limiting the potential costs resulting from spillovers to competitors.

- *Prevent Entry*: The existence of a useful product that is freely available on open-source platforms can increase the difficulty for new competitors to enter the market. Firms

> may therefore choose to make products open source to reduce the expected profitability of entry for potential competitors. This is particularly effective when the firm has competitive advantages in related markets or complementary products. (Bonaccorsi et al., 2006)

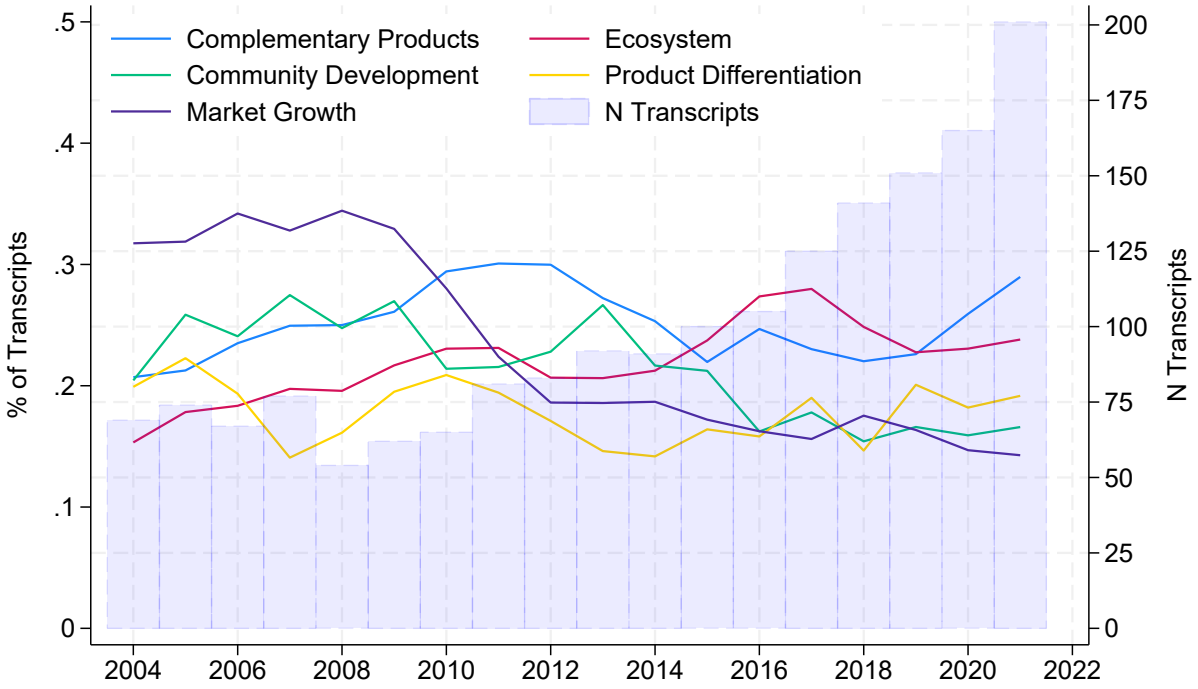
Trends in Discussions of Open Source in Earnings Calls

This section reports time-series trends in discussions of private value generated through open source in earnings calls. These statistics are plotted in Figure IA1. First, we plot the total number of transcripts each year with discussions of open-source. Discussions are relatively constant from 2004 through 2012, then steadily increase through 2016. Thereafter, discussions increase significantly, with more than 200 earnings-call transcripts discussing open source in 2021.

Second, we plot the three-year moving average of the percent of transcripts discussing private value that align with each of the five most-discussed channels. These channels include Complementary Products, Ecosystem, Product Differentiation, Community Development, and Market Growth. The most notable change is for discussions of open-source activities helping grow the market at-large. While Market Growth is the most discussed channel prior to 2009, present in more than 30% of transcripts, it declines precipitously through 2012, present in less than 15% of transcripts by 2021. We also find marginal increases in discussions of complementary products and developing ecosystems during the sample period, such that these are the most discussed channels by 2021.

Figure IA1
Trends in discussions of open source in earnings calls

This figure illustrates time-series trends in managers' discussions of the private value generated from open-source activities in earnings calls. The lines represent three-year moving averages of the percent of transcripts discussing private value that align with each of the five most-discussed channels in the sample. The bars represent the total number of transcripts each year with discussions of open source.



Internet Appendix B

Open-Source Licenses

Table IA1
License Classification Based on Permission Levels

This table classifies common licenses of GitHub repositories based on their permission levels. Permissive licenses place few limitations on the use of source code, while restrictive licenses constrain how the code may be redistributed or integrated with proprietary software. We adopt the Open Source Initiative's definition of open source, which stipulates that an open source license must not discriminate against any person, restrict other software, or be specific to a product, among other criteria. See https://opensource.org/osd/ for details.

Type of licenses	Restrictions	Benefits	Costs	Examples
Permissive License				
Open source and permissive	Keep copyright information	Compatibility with both open source and proprietary projects	Limited protection of the original developer's work	MIT, Apache 2.0
Restrictive License				
Open source and weak copyleft	Copyleft for the original codes	Encourages contributions to the open source component while permitting proprietary integration	Incompatible with proprietary projects with static linking	LGPL
Open source and strong copyleft	Can use but derivative work must also be open source	Preserve the open nature; Monetize	Incompatible with proprietary projects and compliance burden	GPL, AGPL
Source available but limit use in certain products	Limits the use of the software in specific commercial or competitive scenarios	Monetize software by offering additional commercial licenses/through complementary products	Less contribution from both individual users and commercial users	Amazon Software License
Source available but limit commercial use	Only for non-commercial purposes	Encourages contributions and community engagement while protecting against certain commercial uses	Close monitoring and compliance enforcement; Less contribution from commercial users	CC-BY-NC-4.0
No license	By default illegal to use, distribute, or modify the code	IP protection, flexibility in choosing license later	Discourage adoption and contribution	

Internet Appendix C

This appendix details our usage of large language models. We describe the steps taken to classify or evaluate GitHub repositories based on their topics, the extent to which they complement firms' commercial offerings, and their novelty. We also include prompts used to classify private-value-generating channels discussed in earnings calls related to open-source projects.

Classifying GitHub Repositories by Topic

While repository admins can add topics to increase the visibility of their projects, there are no specific requirements regarding which topics they can use. These topic labels can be assigned based on the intended purpose, subject area, affinity groups, or other important qualities. Therefore, the potential choice of topics is unlimited. In addition, many repositories remain unlabeled. To address these challenges, we rely on the GitRanking taxonomy, proposed by Sas et al. (2023), to limit the topic space for repository classification based on a structured set of topics. Additionally, we employ a large language model to infer appropriate topics for repositories by analyzing the content of the repositories and identifying relevant themes, even when explicit topics are not provided.

GitRanking is a taxonomy consisting of 301 labels derived from 121,000 GitHub topics. These 301 topic labels are organized into distinct levels based on their meanings. To balance the breadth and specificity of topics for our purposes, we elect to use the third level, which is comprised of 62 topics. To further refine these topics, we use ChatGPT to group similar topics, enhance clarity, and reduce overlap between topics. This process resulted in 17 distinct topics, which we use to classify the repositories in our sample. Definitions of these topics are provided in the prompt below.

Next, we use ChatGPT to assess how closely a repository aligns with a given topic, assigning a relatedness score ranging from 0 to 1. The information provided to ChatGPT in-

cludes the repository's name, description, main programming language, self-reported topics, and website, all obtained via the GitHub API. We explicitly allow ChatGPT to incorporate external knowledge beyond the provided data. Furthermore, to ensure consistency in evaluating a repository's relatedness to a topic, we developed a scoring reference and repeatedly tested it on a small subsample of repositories, confirming that score variations typically stay within 0.1.

Specifically, we use the following model parameters and prompt for OpenAI's API:

`Model`: gpt-4o-2024-08-06

`Temperature`: 0

`Seed`: 2024

`Prompt`: Assign relevance weights (0 to 1) to predefined topic labels for a GitHub repository. The weight represents how relevant each topic is to the repository. Your weights should follow the scoring reference and definitions of topic labels as below.

Scoring Reference:

- **0.0**: The topic is **completely irrelevant** to the repository. There is **no code, documentation, or features** associated with this topic. The topic does not apply in any way.

- **0.1**: The topic has **extremely low relevance**. It is **mentioned only once** or in a **minor part** of the repository (e.g., a reference in a single file or a brief mention in documentation). There is **no meaningful functionality** related to this topic.

- **0.2**: The topic has **very low relevance**. There is a **small, secondary feature** related to the topic, but it plays a **minimal role** in the repository. It is **not central** to the repository's purpose and may only be used in a **supporting** or **optional capacity**.

- **0.3**: The topic has **low relevance**. The repository includes **some functionality** or content related to the topic, but it is a **minor or auxiliary component**. The topic

is **not integral** to the primary goals of the repository, though it may be referenced or used in **specific parts** of the project.

- **0.4**: The topic has **below-average relevance**. It plays a **noticeable role** in the repository, but it is **not essential**. There are **clear references, code, or features** related to the topic, but they are **not a major focus**.
- **0.5**: The topic has **moderate relevance**. It is one of **several key areas** covered by the repository. A **substantial portion** of the repository's code, features, or documentation relates to this topic, but it is **not the main focus** of the project.
- **0.6**: The topic has **moderately high relevance**. The topic is one of the **core areas** of the repository, with a **significant portion** of the code, features, or documentation focused on it. The repository relies heavily on this topic, but other topics are also important.
- **0.7**: The topic has **high relevance**. A **large portion** of the repository is built around this topic. The topic plays a **central role** in the repository's features, functionality, or design. Most of the repository's content is directly related to this topic, but there are still other areas of focus.
- **0.8**: The topic has **very high relevance**. It is a **primary focus** of the repository, with **most features, code, and documentation** centered around it. Nearly all content is related to this topic, with only a few secondary areas.
- **0.9**: The topic has **near-perfect relevance**. The repository is **almost entirely centered** around this topic. The vast majority of its code, design, and purpose are related to this field, with **only minor mentions** of other topics.
- **1.0**: The topic has **perfect relevance**. The repository's **sole purpose** is to serve this topic. **Every feature, line of code, and piece of documentation** is directly related to this topic, with **no other topics** playing a significant role.

Topic labels to assign weights to:

- **Digital Media**: Projects related to game development, video games, animation, camera software, image processing, audio or video editing, and interactive media.
- **General Data Handling**: Projects focused on managing data, including data structures, file systems, databases, and ETL (Extract, Transform, Load) processes.
- **Advanced Data Analysis**: Projects involving complex data analysis, such as big data, bioinformatics, data science, text analysis, time series analysis, or data visualization.
- **Security**: Projects related to cryptography, data protection, authentication, network security, privacy, or detecting threats like phishing.
- **Cloud Infrastructure and DevOps**: Projects centered on cloud computing, CI/CD pipelines, distributed computing, microservices, backups, infrastructure automation, or serverless computing.
- **Software Engineering**: Projects focused on software architecture, testing, program analysis, design patterns, or software development methodologies like Agile or Scrum.
- **Front-End Web Development**: Projects involving client-side development, including UI/UX design, HTML/CSS, and JavaScript frameworks.
- **Back-End Web Development**: Projects centered on server-side development, including APIs, databases, and routing.
- **Core AI/ML**: Projects focused on foundational machine learning concepts such as neural networks, deep learning, semi-supervised learning, or reinforcement learning.
- **AI Applications**: Projects applying AI techniques in specific domains, such as robotics, computational biology, computer vision, or natural language processing (NLP).

- **Development Tools**: Projects related to building or maintaining programming tools like compilers, interpreters, validators, debugging tools, IDEs, or version control systems.
- **Operating Systems and Platforms**: Projects focused on OS development, kernel development, embedded systems, or platform-specific development (e.g., Windows, Linux, Android).
- **Documentation**: Projects primarily providing manuals, guides, API documentation, or technical standards.
- **Community and Governance**: Projects focused on open-source community guidelines, such as codes of conduct, contribution guidelines, or governance policies.
- **Education and Learning**: Projects designed for educational purposes, such as tutorials, training modules, or coding bootcamps.
- **Configuration and Templates**: Projects offering pre-configured setups, boilerplate code, or templates for quick project initialization (e.g., Dockerfiles, CI/CD configs).
- **Other**: For any project that doesn't fit the above categories. The weight should be zero if no miscellaneous topics are relevant to the repository, or a non-zero value if there are other significant topics.

Use both the repository details and relevant knowledge you have about this repository to make your decision.

Evaluating Complementarity of GitHub Repositories

Similarly, we use a large language model to evaluate the complementarity of GitHub repositories to their owners' commercial products, which is defined as directly supporting or enhancing the firm's core business products. In addition to the repository's name, description, main programming language, self-reported topics, and website, we also include

the Compustat name of the firm owning the repository. We explicitly allow ChatGPT to incorporate external knowledge beyond the provided data, and include a scoring reference. We have ChatGPT not only provide the complementarity score but also the name of the commercial product that this repository complements, with which we confirm the validity of the scores.

To illustrate the intuition of the resulting complementarity scores, we provide the following examples. The first example is the repository "WhatsApp/StringPacks," which is a library designed to store translation strings in a more efficient binary format for Android applications. This library can operate completely independently of WhatsApp's messaging services and be used by any relevant Android application. This repository has a complementary score of 0. In contrast, the second example is the repository "WhatsApp/stickers," which contains iOS and Android sample apps as well as an API for creating third-party sticker packs for WhatsApp. This project directly complements the WhatsApp messaging app, resulting in a complementary score of 0.8.

We use the following model parameters and prompt for OpenAI's API:

`Model`: gpt-4o-2024-08-06

`Temperature`: 0

`Seed`: 2024

`Prompt`:

Evaluate whether a GitHub repository owned by a US public firm complements the firm's commercial products or operates as a standalone project based on the following scoring reference. Use both the repository details and relevant knowledge you have about this repository to make your decision. Only return in JSON and do not include anything else. In your JSON response, include the following fields: `repo_id`, `comp_score` (complementarity score), and `comm_product` (the commercial product to which the repository complements).

Scoring Reference:

- **0.0**: The repository is entirely standalone. It has no overlap or connection with any

of the firm's commercial products. It operates independently of the company's core business.

- **0.1**: The repository shows minimal potential relevance or use in conjunction with the firm's commercial products but is not designed for or marketed as part of the firm's offerings.
- **0.2**: The repository may have slight overlaps with the firm's commercial offerings but is not positioned as a key integration. It could potentially be used with the firm's products but has no direct integration or clear marketing as a complementary tool.
- **0.3**: The repository shows some potential to complement the firm's products but is still largely standalone. There might be some integrations, but they are not essential or exclusive to the firm's ecosystem.
- **0.4**: The repository provides a small but noticeable enhancement to the firm's commercial products. However, the connection to the commercial offering is weak, and the repository is still usable independently.
- **0.5**: The repository offers some clear value to the firm's commercial products but is not a core or exclusive component. It may integrate with or enhance the firm's product, but its relevance is moderate.
- **0.6**: The repository offers strong support for the firm's products and likely exists to enhance or complement the product experience. However, it is still not fully dependent on the commercial product.
- **0.7**: The repository is closely linked with the firm's commercial product and provides substantial enhancements or integrations. It is marketed or documented as a useful component for customers of the firm's product.

- **0.8**: The repository strongly complements the firm's commercial product and is used almost exclusively within the context of that product. However, it may still be used in other contexts with significant effort.

- **0.9**: The repository is nearly indispensable for customers using the firm's commercial product. It is closely tied to the product, and its functionality is largely dependent on it.

- **1.0**: The repository is entirely and exclusively built to complement and support the firm's commercial product. It has no utility outside of the firm's product and is crucial for its full use. The repository cannot function independently and exists solely to enhance the firm's commercial offering.

Category-Specific Definitions:

- **Complementary Repositories**: These repositories directly support or enhance the firm's core business offerings.

- **Standalone Repositories**: These are open-source projects, tools, or experiments that do not contribute to or enhance the firm's main commercial products, even if created or maintained by the firm.

Evaluating Novelty of GitHub Repositories

Lastly, we use a large language model to evaluate the novelty of GitHub Repositories. To do so, we provide the repository's name, description, main programming language, self-reported topics, and website. We explicitly allow ChatGPT to incorporate external knowledge beyond the provided data, and include a scoring reference.

Again, to illustrate the intuition for the resulting novelty scores, we consider the following examples. First, the average novelty score for repositories affiliated with the "LinkedInLearning" organization account is 0.1. These projects are often exercises associated with courses on

the LinkedIn Learning platform. In contrast, the repository "google-deepmind/alphafold", which hosts the open-source code of AlphaFold (an AI system developed by the 2024 Chemistry Nobel Prize winners that predicts a protein's 3D structure), has a novelty score of 0.8, the highest in our sample.[IA1]

We use the following model parameters and prompt for OpenAI's API:

`Model`: gpt-4o-2024-08-06

`Temperature`: 0

`Seed`: 2024

`Prompt`:

Evaluate the originality of a GitHub repository. Originality measures how novel or groundbreaking a repository is compared to existing solutions, focusing on whether it introduces new ideas, techniques, or approaches.

Use both the repository details and relevant knowledge you have about this repository to make your decision. Use the scoring reference provided for consistency. Only return in JSON and do not include anything else.

Originality Scoring Reference:

- **0.0**: The repository introduces no new ideas or techniques. It is a near-complete replication of existing solutions with no modifications.
- **0.1**: The repository shows minimal originality, with minor tweaks or adaptations of well-established methods, but still follows existing patterns closely.
- **0.2**: Low originality. The repository includes slight variations or small improvements on existing solutions but doesn't introduce any novel concepts.
- **0.3**: The repository adds some new ideas or features, but these are incremental and build directly on existing work.

[IA1] Note that the Nobel Prize announcement occurred after the model's training period, and therefore the score was not influenced by the Nobel news.

- **0.4**: Below-average originality. The repository offers a combination of existing techniques with minor innovations or optimizations.
- **0.5**: Moderate originality. The repository introduces some interesting and unique features or techniques, but they are not highly groundbreaking.
- **0.6**: The repository shows notable originality. It provides a fresh approach to solving a problem, though the idea may not be entirely new.
- **0.7**: High originality. The repository demonstrates a new concept or method that has the potential to influence other projects or domains.
- **0.8**: Very high originality. The repository presents a significantly novel approach, introducing new methodologies, techniques, or tools that are not widely available.
- **0.9**: Nearly groundbreaking. The repository offers a unique solution that stands out as highly innovative compared to others in the field.
- **1.0**: Completely original. The repository introduces entirely new concepts, techniques, or tools that could redefine the domain or set new standards.

Classifying Dimensions of Open-Source Engagement in Earnings Calls

You are provided with an excerpt from an earnings call transcript, which includes discussions potentially related to open-source engagement. The excerpt has been selected based on the presence of keywords such as *open source*, *open-source*, *GitHub*, *GitLab*, *open-core*, and *open core*.

Your task is to carefully read the excerpt and determine whether it addresses any of the following three dimensions of open-source engagement:

Task instructions

Identify the types of open-source engagement mentioned. For each dimension below, return 1 if the excerpt clearly discusses it, and 0 otherwise:

1. **Open-source release**: Does the speaker mention the firm releasing its own open-source product or contributing to open-source projects?

2. **Open-source adoption**: Does the speaker mention the firm using or integrating open-source software developed by others?

3. **Open-source competition**: Does the speaker refer to open-source products or communities as competitors that could affect the firm's proprietary offerings or market position?

Output format

Return your answer strictly in the following JSON format. Do not include any extra text or commentary.

```
{
  "file_name": the input file name,
  "release": 1 or 0,
  "adoption": 1 or 0,
  "competition": 1 or 0
}
```

Example Input

```
{
  "file_name": "111.txt",
  "content": "
Question & Answer


Jonathan Goldberg (Deutsche Bank - Analyst)
```

And then are you seeing any impact from some of these open-source map data providers that are emerging?

Harold Goddijn (Tomtom N.V. - CEO)

No, we don't see that. What we have seen in the past, of course, is price pressure on our Licensing deals. But what we have not seen is that we've been replaced by open-source products. It depends a little bit on the type of application, obviously. The quality requirements for Internet display are different than the quality requirements for Automotive turn-by-turn navigation.
So the high-quality map that we offer, we can -- we have some pricing power, of course, in the turn-by-turn navigation space, but of course the attributes are needed and a basic requirement."
}

Example output


```
{
  "file_name": "111.txt",
  "release": 0,
  "adoption": 0,
  "competition": 1
}
```

Labeling Private-Value-Generating Channels Discussed in Earnings Calls

You are provided with an excerpt from a firm's earnings call transcript. This excerpt may discuss the firm's release of open-source products, the firm's contributions to open-source projects, or engagement in open source by other parties.

Your task is to determine whether the discussion reveals information about how **private value** is generated from these open-source activities. If it does, focus specifically on the discussion of open source. Determine whether the discussion aligns with **specific channels** through which private value is generated. The objective is to understand how firms think **private value** is generated from open-source activities. For each value channel, select it only if the excerpt links open-source activity to that specific form of private value. Do not infer likely outcomes. Your assessment should reflect what is said in the excerpt as closely as possible and not your own consideration of how private value may be generated.

Definitions

Private value refers to the value generated for the firm that releases or contributes to open-source projects. This includes value from both the intrinsic merits of the product and from making it open-source.

It is different from **public value**, which benefits other firms or the ecosystem at large.

Task instructions

1. **Determine private value relevance**

Decide if the excerpt discusses the **private value** generated from open-source involvement.

- Return 1 if yes, 0 if not.

2. **Identify relevant value channels**

Assess which of the following channels are mentioned:

- **Network effects**: A firm benefits when its open-source project becomes more valuable as more people use it. Open sourcing accelerates adoption, which increases the utility of the project for all users, including the firm.
- **Community development**: A firm benefits from others contributing code, features, bug fixes, or ideas. This improves the project's quality, lowers internal development and maintenance costs, or enhances organizational learning.
- **Technology control**: A firm uses open source to influence the direction, architecture, or standards of a technology. Adoption by others allows the firm to help shape future development in ways that benefit them.
- **Ecosystem**: An open-source project is part of a firm's ecosystem. The open-source nature of the project promotes adoption of the firm's ecosystem, increasing switching costs for users and making it less likely they switch to the firm's competitors.
- **Complementary products**: A firm's open-source project supports, enhances, or drives demand for at least one of its separate proprietary products.
- **Growing the market**: Open-source activity contributes to market growth. A firm then benefits as a participant in this larger market, even if the benefits are also shared by others.
- **Employee reputation**: A firm allows or encourages employees to contribute to open source to boost their visibility and reputation in the open-source community. This may be seen as a perk for valuable employees who care about their reputation in the open-source community.
- **Identify talent**: Open-source engagement allows a firm to observe and identify external contributors who may become high-quality hires.

- **Integration costs**: Because an open-source project is public, new hires already familiar with it can be onboarded faster, reducing training or integration time.
- **Reputation**: Engagement in open source may increase a firm's reputation among developers or customers as being community-oriented.
- **Transparency**: A firm highlights that open source provides visibility into the code or development process, which is valued by customers or developers.
- **Quality certification**: A firm uses open source as a signal of quality or reliability because public review by the community validates the project.
- **Undermine competitors**: A firm uses open source to weaken or displace a competitor's proprietary product by offering a free or open alternative.
- **Prevent entry**: A firm releases open-source projects to make it more difficult for new firms to enter into that product market.
- **Signaling innovation direction**: A firm releases open-source projects to signal their intention to move into a technological space they have not previously operated in.
- **Inherent excludability**: A firm shares a project via open source that, despite being public, is difficult for competitors to use, reducing concerns about the costs of sharing projects via open source.
- **Product differentiation**: Open source is a characteristic that can help differentiate a firm's products from its competitors.

Output Format

Return only the following JSON. Do not include any explanation or extra content.

```
{"file_name": the input file name,
  "private_value": 1 or 0,
```

```
  "channels": {
    "network_effects": 1 or 0,
    "community_dev": 1 or 0,
    "tech_control": 1 or 0,
    "ecosystem": 1 or 0,
    "comp_products": 1 or 0,
    "market_growth": 1 or 0,
    "emp_reputation": 1 or 0,
    "id_talent": 1 or 0,
    "integration_costs": 1 or 0,
    "reputation": 1 or 0,
    "transparency": 1 or 0,
    "quality": 1 or 0,
    "undermine_competitors": 1 or 0,
    "prevent_entry": 1 or 0,
    "signaling": 1 or 0,
    "inherent_excludability": 1 or 0,
    "product_diff": 1 or 0
  }
}
```

Example Input

```
{"file_name": "111.txt",
"content": "
Xueji Wang (Tuya Inc. - Founder, CEO & Director)
```

[Interpreted] Let me address the question by first clarifying what Matter is. It is an IoT protocol formed by integrating the technical characteristics of home kits, open [SaaS] and Zigbee 3.0.

The integrated protocol focuses on the field of smart homes and the local interconnections. Having multiple influential enterprises developing the protocol together in an open source format will accelerate its adoption, which, in turn, will help improve the low penetration rate of IoT in home appliances.

...

In an article about the progress of Matter published by the Wall Street Journal on February 22, the only app demo displayed was Tuya's, which not only showcase that our extensive influence in IoT PaaS field, but also highlighted the synergies between Tuya and Matter."

}

Example Output


```
{"file_name": "111.txt",
  "private_value": 1,
  "channels": {
    "network_effects": 1,
    "community_dev": 0,
    "tech_control": 1,
    "ecosystem": 1,
    "comp_products": 0,
    "market_growth": 1,
    "emp_reputation": 0,
    "id_talent": 0,
```

```
    "integration_costs": 0,
    "reputation": 1,
    "transparency": 0,
    "quality": 0,
    "undermine_competitors": 0,
    "prevent_entry": 0,
    "signaling": 0,
    "inherent_excludability": 0,
    "product_diff": 1
  }
}
```

Internet Appendix D

This appendix presents additional analyses of open-source activity and repository private value. First, we investigate the correlations between firm characteristics and open-source activity in a regression setting. Second, we investigate the strategic timing of repository releases. Third, we present results from alternative specifications for the validity tests from Section 4.3.1. Fourth, we investigate the correlations between firm characteristics and repository private value in a regression setting.

Determinants of Open-Source Activity

To provide a more rigorous characterization of open-source firms, this section considers the determinants of open-source activity in a regression setting. This approach controls for time-varying industry and time-invariant firm fixed effects. It also allows us to test the relative strength of the correlation between variables and open-source activity to identify major determinants of open-source activity. These tests are intended to be descriptive rather than causal, helping researchers understand potential selection bias in open-source activity and identify relevant omitted variables in a given research context.

The results from this analysis are reported in Table IA2. The data pertains to firm-month observations of all public firms. Each regression includes the full set of firm and product market characteristics discussed in the previous section, as well as industry-time fixed effects. Columns (2) and (4) report regressions that also include firm fixed effects. Standard errors are double clustered on industry and time, and all independent variables are standardized to facilitate interpretation.

The regressions reported in Columns (1) and (2) pertain to all public firms. The dependent variable is an indicator that equals one if the firm is an open-source firm.[IA2] The results in Column (1) therefore reflect the average difference between open-source and non-

[IA2] Note that prior to the firm's first open-source activity, it is classified as a non-open-source firm.

open-source firms within each industry. We find that firms with higher valuations (Mkt Cap, Market-to-Book) and more innovation (N Patents, R&D Exp) are more likely to be open-source firms. However, these firms also appear to be less profitable than their industry peers (Return-on-Assets) and have lower annual returns. One possible interpretation is that open-source firms may follow a traditional strategy in technology industries of keeping profit margins low in the short term to maximize future growth.

Interestingly, almost all of these relations appear to be a function of stable firm differences. When including firm fixed effects in Column (2), all variables (except annual returns) become statistically insignificant. A firm's average characteristics therefore do not appear to significantly differ from before to after their first open-source activity. In combination with a relatively large adjusted R^2 for this regression, it seems that firm fixed effects can account for most of the differences between open-source and non-open-source firms.

In Columns (3) and (4), we focus only on open-source firms and examine the determinants of monthly open-source activity. To measure open-source activity, we use the number of commits to repositories owned by the firm in that month. Column (3) compares firms to their industry peers and reports similar results as those reported in Column (1). Specifically, firms with more open-source activities tend to be larger and more innovative, although they also tend to be less profitable and have lower returns. These firms also tend to have more market power, but receive fewer benefits from product-market network effects and face a more fluid product market.

However, the relations for product-market characteristics reverse when firm fixed effects are included in Column (4). For example, while firms with more market power tend to have more open-source activities, these firms are especially active when their market power is lower relative to their sample average. This result suggests that firms may use open-source activities to maintain their market power, however further research is required to make a stronger conclusion. We also find that firms engage in more open-source activities when their performance is relatively weak (Sales Growth, Annual Returns) and, intuitively, when they

have more employees.

Strategic Timing of Repository Releases

In this section, we investigate the extent to which firms strategically time repository releases around certain events or following trends in stock prices. This possibility is particularly problematic for our analysis if it results in an overestimation of value for the affected repositories. For events, we consider product releases, software-developer conferences, patent grants, and earnings announcements. For trends in stock prices, we consider the firm's market-adjusted return over the five days prior to the repository release and the overall stock market return over the same period.

We first examine the timing of repository releases. To do so, we construct a daily stock-level panel dataset and define an indicator variable that equals one if the firm releases a repository on that day. We then create separate indicator variables for each of our four events that equal one if the day overlaps with the event.

To identify product release dates, we use data on trademarks from the USPTO. A product must be in commercial use before a trademark is granted, so the data on trademark applications include the first date that the potential trademark was used in a commercial capacity. We use these dates to proxy for product release dates. We match trademark data to our sample of GitHub firms using a fuzzy name-matching algorithm, which identifies 34,998 trademarks granted to firms with GitHub repositories from 2015 through 2023. To identify software-developer conference dates, we obtain the dates, from 2015 through 2023, for eight major conferences: GitHub Universe, Google I/O, Microsoft Build, Apple WWDC, AWS re:Invent, FOSDEM, KubeCon + CloudNativeCon, and PyCon. To identify patent grant dates, we use data provided by Kogan et al. (2017). To identify earnings announcement dates, we use data from IBES.

We then regress the indicator for a repository being released, expressed as a percent, on the indicators for our various events, along with the five-day market-adjusted return and

five-day overall market return. We also control for market capitalization on the day before the repository release. We include day-of-week fixed effects to control for different release rates across days of the week,[IA3] firm-year fixed effects to control for annual-varying firm characteristics, and double-clustered standard errors by firm and year.

The results from variations of this regression are presented in Panel A of Table IA3. Column (1) reports results from a linear probability model using the full sample of daily stock data. We find that repositories are more likely to be released on the same day as a product release and during developer conferences. However, there is no significant overlap with patent grant dates and earnings announcement dates, nor is there a significant relation with firm- or market-level trends in stock prices. Column (2) limits the sample to firms that release a repository in our sample, and Column (3) uses a probit model. In both cases, we continue to find qualitatively similar results.

Thus, we find that firms may time repository releases around product releases and developer conferences. This strategic timing can distort our analysis if this introduces an upward bias in repository value. To investigate this, we regress repository private value on the same independent variables, following the specification in Column (3) of Table 4. We find that repositories released along with products and during developer conferences are estimated to be significantly less valuable. Thus, these repositories are, if anything, under-weighted in our analysis. Moreover, these results are not a function a firm-specific characteristics, as the relations remain statistically significant when including firm-year fixed effects (untabulated).

In Panel B of Table IA3, we investigate the relation between private value and future popularity for repositories that are released during one of these firm events. The regression specification follows that of Column (3) of Table 4. Columns (1) through (4) limit the sample to repositories releases overlapping with product releases, developer conferences, patent grants, and earnings announcements, respectively. Column (5) combines each of these samples. In total, 8,042 out of a total of 28,090 repositories overlap with at least one

[IA3] We find that repository releases are monotonically decreasing across days of the week, and patents are granted only on Tuesdays.

firm event.

Across all five regressions, we find a generally weaker relation between private value and future popularity compared to Column(3) of Table 4, consistent with these events diluting the relevance of stock returns for repositories. Nonetheless, the relation is positive for all four samples, and is statistically significant for the collection of all events (Column (5)). Thus, there still appears to be some information about the repository contained in investor reactions to these repository releases.

Validation: Alternative Specifications

This section reports alternative specifications for the validation tests of the private value of open-source innovation. For each alternative specification, we regress the measure of private value on future repository popularity (i.e., stars) following the same specification as in Column (3) of Table 4, including industry-year fixed effects and double clustering standard errors on firm and year.

The first set of results are reported in Table IA4. In Column (1), we use the number of forks as of February 2024 as an alternative measure of future popularity. We continue to find a strong relation between repository value and future popularity using this measure.

In Column (2), we exclude repositories from Amazon Inc., Microsoft Corp., and Alphabet Inc. from the regression to ensure our results are not driven by the three firms with the most repositories in our sample. After excluding these repositories, we still find a statistically significant relation between private value (i.e., ξ) and future repository popularity. In fact, the economic magnitude of the relation is slightly larger than that reported for our full sample in Column (3) of Table 4. Thus, the value-relevant information contained in our estimates is not concentrated in repositories released by these three firms.

In Column (3) of Table IA4, we replace the dependent variable with an estimate of ξ using returns on only the release day, t, instead of days t through $t+2$. We continue to find a statistically significant relation between private value and future repository popu-

larity, with the economic magnitude being almost identical to that reported for our full sample in Column (3) of Table 4. Thus, our results appear to be primarily a function of returns on the release day, which is consistent with our finding in Figure 3 that abnormal trading value is significantly positive on the release day. Nonetheless, we report our main results using the three-day window to ensure comparability to prior studies using a similar methodology (Kogan et al., 2017; Desai et al., 2025).

In Column (4) of Table IA4, we replace the dependent variable with ξ^{alt}, the three-day cumulative market-adjusted release return (R) multiplied by market capitalization (M), which we evenly divide among repositories released on the same day (N). Given that this variable is highly skewed and takes both positive and negative values, we apply an inverse hyperbolic sine transformation. We find that the relation between this measure of release returns and future repository popularity is also statistically significant. Thus, our results are not solely a function of the distributional assumptions made when converting R to ξ, particularly that repositories have strictly positive values.

Determinants of Repository Value: Firm Characteristics

In Table IA5, we investigate firm characteristics. These characteristics include market-to-book ratio, return-on-assets, investment, annual stock return, annual sales growth, tangibility, and R&D expenditure. For observations with missing R&D expenditure, we replace the value with zero and set an indicator variable, R&D Exp Missing, equal to one. We find that none of these variables are significantly related to repository value in the cross section. The lack of statistical significance across firm characteristics after including our standard controls (market capitalization, volatility, number of employees, and patent value), particularly in contrast to the results for repository and product market characteristics, suggests these controls capture much of the firm-level variation in repository value. This finding narrows the scope for potentially omitted variables that could confound our analysis of firm growth in Section 5.

Table IA2
Determinants of Open-Source Activity

This table reports regression results to examine the factors influencing the extensive and intensive margins of open-source activity on GitHub among U.S. public firms. The dependent variable in Columns (1) and (2), *Github*, is an indicator variable that equals one from the first month a firm releases a repository on GitHub. In Columns (3) and (4), the dependent variable is the natural logarithm of one plus the number of commits made to repositories owned by the firm each month. See Table IA7 for the definition of variables. Standard errors double clustered by firm and year are reported in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

	(1) GitHub	(2) GitHub	(3) ln(N Commits + 1)	(4) ln(N Commits + 1)
ln(Mkt Cap)	0.081***	0.013	1.230***	0.101
	(0.010)	(0.010)	(0.227)	(0.219)
ln(Employees)	0.008	0.024	-0.031	0.406
	(0.011)	(0.017)	(0.204)	(0.276)
ln(N Patents + 1)	0.050***	0.039	0.501***	-0.147
	(0.008)	(0.027)	(0.177)	(0.429)
Market-to-Book	0.012**	0.000	-0.001	0.088
	(0.006)	(0.004)	(0.101)	(0.061)
Return-on-Assets	-0.007	0.003	-0.153*	-0.003
	(0.004)	(0.003)	(0.084)	(0.048)
Investment	0.003	-0.000	0.024	0.134**
	(0.005)	(0.002)	(0.118)	(0.054)
Return (t-12 to t-1)	-0.006***	-0.002**	-0.074*	-0.054**
	(0.002)	(0.001)	(0.043)	(0.024)
Sales Growth	-0.000	-0.001	0.024	-0.074*
	(0.003)	(0.002)	(0.059)	(0.041)
Tangibility	-0.002	-0.018**	0.093	0.005
	(0.010)	(0.009)	(0.142)	(0.131)
R&D Exp/Total Assets	0.029***	-0.003	0.342**	0.084
	(0.008)	(0.007)	(0.144)	(0.081)
R&D Exp Missing	-0.032**	-0.020	0.212	-0.598
	(0.016)	(0.020)	(0.317)	(0.393)
Market Power	0.008	0.001	0.220*	-0.087
	(0.006)	(0.003)	(0.123)	(0.081)
Product Market Centrality	-0.046***	-0.008	-0.604***	0.156
	(0.015)	(0.013)	(0.154)	(0.144)
Scope	-0.007	0.004	0.037	-0.083
	(0.007)	(0.006)	(0.096)	(0.066)
Product Market Similarity	0.009	-0.004	0.245**	-0.089
	(0.010)	(0.008)	(0.114)	(0.075)
Product Market Fluidity	0.014**	-0.005	0.132	-0.062
	(0.006)	(0.003)	(0.095)	(0.058)
Observations	199,961	199,948	25,923	25,914
Adj. R^2	0.346	0.871	0.331	0.784
Industry x Time FE	✓	✓	✓	✓
Firm FE		✓		✓
Sample	All firms	All firms	GitHub = 1	GitHub = 1

Table IA3
Strategic Timing of Repository Releases

This table investigates the strategic timing of repository releases. Columns (1) through (3) of Panel A examine a daily firm-level panel dataset, while Column (4) of Panel A and all columns of Panel B examine a repository-level dataset. All columns of both panels report OLS regressions except Column (3) of Panel A, which reports a probit regression. The dependent variable in Columns (1) through (3) of Panel A is an indicator variable that equals one if a repository is released on that day. The dependent variable in Column (4) of Panel A and all columns of Panel B is the repository private value (ξ). Each column of Panel B examines the set of repositories whose release dates overlap with a given event: trademark first use, developer conference, patent grant, earnings announcement, or any of those four events. Each variable is defined, along with its data source, in Appendix IA7. Standard errors double clustered by firm and year are reported in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

Panel A: Repository timing and value

	(1) Repo Posted	(2) Repo Posted	(3) Repo Posted	(4) ln(ξ)
Trademark First Use	0.577**	0.763**	0.130***	-0.341**
	(0.212)	(0.278)	(0.039)	(0.118)
Developer Conference	0.017*	0.177*	0.090***	-0.212***
	(0.008)	(0.078)	(0.035)	(0.052)
Patent Grant	0.046	0.082	-0.026	0.039
	(0.027)	(0.083)	(0.050)	(0.026)
Earnings Announcement	-0.013	-0.117	-0.062	0.096
	(0.034)	(0.290)	(0.154)	(0.070)
Return (t-5,t-1)	0.007	0.169	0.143	-1.084
	(0.007)	(0.144)	(0.097)	(0.787)
Market Return (t-5,t-1)	-0.066	-0.713	-0.374	-0.009
	(0.076)	(0.762)	(0.374)	(0.406)
Observations	8,650,989	833,179	396,548	28,090
Adj. R^2	0.308	0.297	0.372	0.828
Controls	✓	✓	✓	✓
Day-of-week FE	✓	✓	✓	
Firm x Year FE	✓	✓	✓	
Repo Topic FE				✓
Industry x Year FE				✓
Regression	OLS	OLS	Probit	OLS
Sample	All firms	GitHub = 1	GitHub = 1	All repos

Panel B: Firm events

	(1)	(2)	(3)	(4)	(5)
ln(Stars + 1)	0.017	0.026	0.021**	0.033	0.031**
	(0.026)	(0.020)	(0.008)	(0.019)	(0.012)
ln(Mkt Cap)	0.824**	0.772***	0.723***	0.750***	0.707***
	(0.299)	(0.069)	(0.061)	(0.034)	(0.058)
ln(Volatility)	1.331**	1.530***	1.361***	1.163***	1.423***
	(0.483)	(0.119)	(0.180)	(0.280)	(0.135)
ln(Employees)	0.105	0.077	0.146	0.104	0.093
	(0.239)	(0.072)	(0.080)	(0.074)	(0.073)
ln(Patent Value + 1)	0.061	0.026	-0.045	-0.001	0.056*
	(0.124)	(0.031)	(0.058)	(0.036)	(0.028)
Observations	748	3,310	4,831	95	8,042
Adj. R^2	0.704	0.785	0.703	0.930	0.750
Repo Topic FE	✓	✓	✓	✓	✓
Industry x Year FE	✓	✓	✓	✓	✓
Sample	Trademark	Conference	Patent	Earnings	All events

Table IA4
Alternative Specifications for Validation Tests

This table presents alternative specifications for regressions of the private value of repositories on future popularity. Column (1) excludes Amazon.com, Inc., Microsoft Corp., and Alphabet Inc. from the sample. Column (2) uses a measure of ξ estimated with the one-day release return as the dependent variable. Column (3) uses the inverse hyperbolic sine transformation of ξ^{alt} as the dependent variable. Each variable is defined, along with its data source, in Appendix IA7. All independent variables are standardized. Standard errors double clustered by firm and year and are reported in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

	(1)	(2) Ex. AMZN, MSFT, GOOG	(3) Dep. = One-day ξ	(4) Dep. = $\sinh^{-1}(\xi^{alt})$
ln(Forks + 1)	0.038***			
	(0.011)			
ln(Stars + 1)		0.113***	0.084***	0.439*
		(0.024)	(0.018)	(0.227)
ln(Mkt Cap)	1.873***	1.929***	1.838***	-0.563
	(0.116)	(0.061)	(0.108)	(0.509)
ln(Volatility)	0.596***	0.449***	0.583***	-0.028
	(0.054)	(0.036)	(0.056)	(0.343)
ln(Employees)	0.142	0.057*	0.153	0.490
	(0.118)	(0.030)	(0.116)	(0.348)
ln(Patent Value + 1)	0.042	-0.018	0.048	0.314
	(0.037)	(0.034)	(0.037)	(0.411)
Observations	28,095	13,331	28,095	28,095
Adj. R^2	0.824	0.873	0.826	0.013
Repo Topic FE	✓	✓	✓	✓
Industry x Year FE	✓	✓	✓	✓

Table IA5
Determinants of Repository Value: Firm Characteristics

This table reports which firm characteristics are correlated with repository private value (ξ). Control variables include market capitalization, volatility, employees and patent value. See Table A1 for the definition of variables. All independent variables are standardized. Standard errors double clustered by firm and year are reported in parentheses. ***, **, and * indicate significance at the 1, 5, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ln(Stars + 1)	0.085***	0.080***	0.081***	0.085***	0.083***	0.083***	0.081***	0.077**
	(0.018)	(0.019)	(0.018)	(0.018)	(0.018)	(0.019)	(0.018)	(0.023)
Market-to-Book	0.004							-0.004
	(0.035)							(0.031)
Return-on-Assets		0.063						0.069
		(0.062)						(0.060)
Investment			0.046					0.037
			(0.089)					(0.111)
Return (t-1)				0.016				0.019
				(0.025)				(0.020)
Sales Growth					0.044			0.027
					(0.044)			(0.048)
Tangibility						0.028		-0.070
						(0.100)		(0.122)
R&D Exp/Total Assets							0.091	0.089
							(0.063)	(0.054)
R&D Exp Missing							0.178	0.128
							(0.312)	(0.304)
Observations	26,949	26,949	26,949	26,949	26,949	26,949	26,949	26,949
Adj. R^2	0.812	0.813	0.813	0.813	0.813	0.813	0.813	0.814
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Repo Topic FE	✓	✓	✓	✓	✓	✓	✓	✓
Industry x Year FE	✓	✓	✓	✓	✓	✓	✓	✓

Internet Appendix E

Estimating Repository Value

This appendix provides an extended description of our procedure to estimate repository value. Further details and discussion of assumptions can be found in Kogan et al. (2017).

The procedure involves observing stock returns in the three-day window following the release of the repository, $[t, t+2]$. Having a short window is important to limit the probability of other events contaminating the estimates. While expanding the window could capture more of the market reaction to minor repositories that do not induce a significant or immediate reaction from investors upon release, the additional noise could render the estimates largely uninformative. It is also possible that there is some information leakage in the days before the release. We elect to use the window $[t, t+2]$ to ensure the comparability of our estimates with those for other assets (Kogan et al., 2017; Desai et al., 2025). Nonetheless, in untabulated regressions, we find quantitatively similar results for all tests when expanding the window to $[t-2, t+2]$.[IA4]

To begin our estimation procedure, we remove fluctuations in daily returns attributable to market movements by subtracting the market return from each firm's daily return. We then cumulate these market-adjusted returns over the three-day event window for repository i, which we label R_i. We assume that R_i is a function of both investor reaction to the repository release, v_i, and idiosyncratic noise, ε_i, such that

$$R_i = v_i + \varepsilon_i. \tag{4}$$

[IA4] In estimating patent values, Kogan et al. (2017) apply an adjustment for the fact that the events are patent *grants*, while information regarding the patent is first revealed to investors 18 months after the patent application is filed. The market reaction on the grant date therefore reflects only a portion of the value corresponding to the resolution of uncertainty as to whether the patent is granted. In the context of GitHub repositories, however, this adjustment is not needed. Information about repositories is not systematically shared prior to the repositories being open-sourced, and so market reactions within the event windows reflect the full value of the repositories.

We construct the estimate of repository value as the product of the investor reaction to the repository release and the firm's market capitalization on the day prior to the release. If multiple repositories are announced on the same day, we assume the value is evenly distributed across those repositories. Given that repository releases do not follow a typical schedule,[IA5] multiple repository releases on the same day tend to, anecdotally, correspond to a single project. The value of repository i, ξ_i, is thus calculated as

$$\xi_i = \frac{1}{N_i} E[v_i|R_i]M_i, \tag{5}$$

where N_i is the number of repositories announced on that day, $E[v_i|R_i]$ is the expected return attributable to the repository release conditional on observing the three-day cumulative market-adjusted return R_i, and M_i is the market capitalization of the firm on the day prior to the repository release.

To estimate the conditional expected return in Equation (5), we adopt the same distributional assumptions about v and ε as Kogan et al. (2017).[IA6] Note that the distributional assumption regarding v_i implies that repositories have strictly positive values. While it is possible that open-source projects provide value to competitors that make the projects less valuable to the firm itself, we assume that firms will only choose to make projects open source if the net effect still results in a positive value for the firm.

Also note that this assumption implies that it is the net present value of repositories that is strictly positive, not necessarily the release returns. Release returns are net of investors' expectations of the value of the firm's R&D expenditure. The true net present value of the repository is thus the release return plus the expected value of the R&D expenditure spent on the repository. Our assumption is that this quantity is strictly positive. However, given that we cannot observe the expected value of the R&D expenditure spent on a given

[IA5] In comparison, patent grants are announced every Tuesday.

[IA6] Specifically, we assume v_i follows a normal distribution truncated at zero such that $v_i \sim \mathcal{N}^+(0, \sigma^2_{vft})$ and ε_i follows a normal distribution such that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_{\varepsilon ft})$. Thus, both distributions vary across firms, f, and time, t.

repository, this introduces an underestimation bias in our estimate of private value that is inherent to the underlying methodology of estimating the economic value of innovation. At maximum, this bias is equal to the expected value of the associated R&D expenditure for repositories with a positive release return, and approaches zero for repositories with zero true net present value.

Under these assumptions, the conditional expected return can be calculated as

$$E[v_i|R_i] = \delta R_i + \sqrt{\delta}\sigma_{\varepsilon ft}\frac{\phi\left(-\sqrt{\delta}\frac{R_i}{\sigma_{\varepsilon ft}}\right)}{1-\Phi\left(-\sqrt{\delta}\frac{R_i}{\sigma_{\varepsilon ft}}\right)}, \tag{6}$$

where ϕ and Φ represent the standard normal PDF and CDF, respectively, and δ denotes the signal-to-noise ratio,

$$\delta = \frac{\sigma^2_{vft}}{\sigma^2_{vft}+\sigma^2_{\varepsilon ft}}. \tag{7}$$

We adopt the same simplifying assumption as Kogan et al. (2017) that δ is the same for all firms and all time periods. We believe this assumption is reasonable in our setting due to the relatively short time period, which begins in 2015. This assumption still allows σ^2_{vft} and $\sigma^2_{\varepsilon ft}$ to vary across firms and time, but only in constant proportion. To estimate δ, we compare the variance of returns in the event window to that of returns over other three-day periods for the same firm within the same year. This comparison takes the regression form

$$\ln(R^2_{fd}) = \gamma I_{fd} + \lambda_{dow} + \eta_{fy} + u_{fd}, \tag{8}$$

where R_{fd} is the three-day cumulative market-adjusted return for firm f on day d, I_{fd} is an indicator variable that equals one if there is a repository released by firm f on day d, λ_{dow} are day-of-week fixed effects, and η_{fy} are firm-year fixed effects. Importantly, this regression only includes firms that have a repository released at some point in the sample period. The estimated $\hat{\delta}$ can be calculated from the resulting estimate $\hat{\lambda}$ as $\hat{\delta} = 1 - e^{-\hat{\gamma}}$. For our main sample of repositories with available public dates, $\hat{\gamma} = 0.0359$ and $\hat{\delta} = 0.0353$.

Finally, we estimate $\sigma^2_{\varepsilon ft}$ for each firm within each year as

$$\sigma^2_{\varepsilon ft} = \frac{3\sigma^2_{ft}}{1 + 3d_{ft}(e^{-\hat{\gamma}} - 1)}, \tag{9}$$

where d_{ft} is the fraction of days in the given year that are release days for firm f and σ^2_{ft} is the variance of daily market-adjusted returns calculated within each firm for each year.

Internet Appendix F

Correlations

Table IA6
Correlation Matrix

This table presents a correlation matrix of all variables included in our analysis of the determinants of repository private value. Each variable is defined, along with its data source, in Appendix IA7.

	ln(ξ)	ln(Stars + 1)	ln(Mkt Cap)	ln(Volatility)	ln(Employees)	ln(Patent Value + 1)	Restrictive License	Template	Complementarity
ln(Stars + 1)	0.258								
ln(Mkt Cap)	0.855	0.225							
ln(Volatility)	-0.262	-0.188	-0.501						
ln(Employees)	0.689	0.084	0.858	-0.420					
ln(Patent Value + 1)	0.737	0.213	0.875	-0.538	0.773				
Restrictive License	-0.167	-0.152	-0.238	0.129	-0.232	-0.190			
Template	-0.007	-0.007	0.005	0.009	-0.018	-0.003	0.023		
Complementarity	-0.007	-0.023	0.094	-0.037	0.119	0.107	-0.108	0.019	
Novelty	0.204	0.434	0.133	0.038	0.067	0.109	-0.105	-0.041	-0.161
ln(Repo Size + 1)	0.027	0.345	0.030	-0.056	-0.020	0.058	0.017	-0.020	0.114
ln(N Repos + 1)	0.559	0.045	0.693	-0.317	0.584	0.601	-0.181	0.041	0.118
ln(Issues Opened + 1)	0.074	0.729	0.075	-0.172	-0.022	0.103	-0.125	-0.020	0.152
Market-to-Book	0.071	-0.102	0.062	0.216	0.006	-0.056	-0.034	0.007	0.124
Return-on-Assets	0.504	0.198	0.544	-0.374	0.407	0.551	-0.059	-0.028	-0.081
Investment	0.403	0.049	0.495	0.049	0.643	0.356	-0.172	-0.016	0.092
Return (t-12 to t-1)	0.188	0.029	0.163	0.043	-0.018	0.034	-0.017	0.016	0.035
Sales Growth	0.097	0.052	0.021	0.359	-0.046	-0.171	-0.044	0.005	0.029
Tangibility	0.425	-0.010	0.514	0.051	0.720	0.378	-0.174	-0.009	0.075
R&D Exp/Total Assets	0.086	0.009	0.047	0.411	0.110	-0.035	-0.078	-0.022	0.196
R&D Exp Missing	-0.192	-0.082	-0.216	-0.005	-0.100	-0.282	0.042	-0.015	-0.178
Market Power	-0.154	0.093	-0.248	0.115	-0.450	-0.202	0.107	0.007	-0.143
ln(PM Centrality)	-0.178	-0.022	-0.214	-0.002	-0.189	-0.214	0.037	-0.007	-0.014
Scope	-0.063	0.043	-0.064	-0.208	-0.324	-0.001	-0.017	0.051	0.013
PM Similarity	-0.196	-0.015	-0.202	-0.096	-0.368	-0.109	0.026	0.021	-0.016
PM Fluidity	-0.101	0.158	-0.090	-0.295	-0.190	-0.034	-0.045	0.000	0.032

	Novelty	ln(Repo Size + 1)	ln(N Repos + 1)	ln(Issues Opened + 1)	Market-to-Book	Return-on-Assets	Investment	Return (t-12 to t-1)
ln(Repo Size + 1)	0.142							
ln(N Repos + 1)	0.073	0.020						
ln(Issues Opened + 1)	0.253	0.376	-0.016					
Market-to-Book	-0.035	-0.021	0.092	-0.108				
Return-on-Assets	0.173	0.064	0.286	0.092	-0.102			
Investment	0.135	-0.066	0.338	-0.068	0.130	0.176		
Return (t-12 to t-1)	0.037	0.004	0.111	-0.008	0.267	0.011	0.038	
Sales Growth	0.095	-0.047	0.019	-0.030	0.326	-0.129	0.383	0.247
Tangibility	0.116	-0.080	0.358	-0.123	0.148	0.172	0.892	0.018
R&D Exp/Total Assets	0.073	-0.035	0.033	-0.063	0.387	-0.206	0.527	0.164
R&D Exp Missing	-0.034	-0.043	-0.335	-0.055	-0.154	-0.058	-0.151	-0.101
Market Power	0.096	0.058	-0.114	0.108	0.023	0.097	-0.214	-0.088
ln(PM Centrality)	-0.073	-0.007	-0.264	0.027	-0.106	-0.201	-0.329	-0.087
Scope	-0.067	0.092	0.045	0.116	-0.063	0.034	-0.559	0.120
PM Similarity	-0.084	0.061	-0.056	0.090	-0.069	-0.227	-0.494	-0.022
PM Fluidity	-0.078	0.049	-0.147	0.199	-0.191	-0.156	-0.445	-0.019

	Sales Growth	Tangibility	R&D Exp/Total Assets	R&D Exp Missing	Market Power	Scope	ln(PM Centrality)	PM Similarity
Tangibility	0.256							
R&D Exp/Total Assets	0.519	0.488						
R&D Exp Missing	-0.126	-0.076	-0.346					
Market Power	0.129	-0.332	-0.072	-0.111				
ln(PM Centrality)	-0.180	-0.242	-0.068	0.260	-0.030			
Scope	-0.225	-0.577	-0.317	-0.012	0.145	0.282		
PM Similarity	-0.181	-0.558	-0.300	0.017	0.211	0.379	0.595	
PM Fluidity	-0.267	-0.434	-0.186	0.112	-0.002	0.670	0.573	0.475

Internet Appendix G

Table IA7
Internet Appendix Variable Definitions

Variable	Definition
Complementarity	Score between zero and one that measures how much the repository complments the firm's commercial products (ChatGPT).
Developer Conference	Indicator variable that equals one if there is a major software-developer conference on that date. Major conferences include GitHub Universe, Google I/O, Microsoft Build, Apple WWDC, AWS re:Invent, FOSDEM, KubeCon + CloudNativeCon, and PyCon.
Earnings Announcement	Indicator value that equals one if the firm has an earnings announcement on that date (IBES).
Employees	Number of employees (Compustat).
Forks	Number of forks received by a repository as of February 2024 (GitHub API).
GitHub	Indicator variable that equals one after the firm releases its first repository (GHArchive).
Investment	CAPX scaled by lagged total assets (Compustat).
Market Capitalization	Share price times the number of shares outstanding (CRSP).
Market Power	An estimate of markups assuming constant returns to scale developed by (Pellegrino, 2025).
Market-to-Book	Ratio of market capitalization to book equity, where book equity is calculated following Davis et al. (2000) (CRSP, Compustat).
N Commits	Number of commits across all repositories owned by the firm in that month (GHArchive).
N Issues Opened	Cumulative number of issues opened for a repository as of December 31, 2023 (GHArchive).
N Repos (t)	Cumulative number of repositories released by a firm prior to month t (GHArchive).
Novelty	Score between zero and one that measures how novel or groundbreaking a repository is compared to existing solutions, focusing on whether it introduces new ideas, techniques, or approaches (ChatGPT).
Number of Patents	Number of patents granted (Kogan et al., 2017).
Patent Grant	Indicator value that equals one if the firm is granted a patent on that date (Kogan et al., 2017).
Patent Value	An estimate of the economic value of patents using stock market returns around the patent grant date (Kogan et al., 2017)

Continued on the next page

Continued

Variable	Definition
Product Market Centrality	Eigenvector centrality calculated from a network created by product market similarity scores (Hoberg and Phillips, 2016).
Product Market Fluidity	A measure of how intensively the product market around a firm is changes (Hoberg et al., 2014).
Product Market Similarity	A measure of how similar a firm's products are to its peers', from Hoberg and Phillips (2016) (Hoberg-Phillips Data Library).
R&D Exp/Total Assets	Research and development expense scaled by lagged total assets (Compustat).
R&D Exp Missing	Indicator variable equal to one if R&D expense is missing (Compustat).
Repo Size	Byte size of a repository as of February 2024 (GitHub API).
Repo Posted	Indicator variable equal to one if a repository is posted on that date (GHArchive).
Restrictive License	Indicator variable that equals one if the repository has a license that restricts commercial use (GitHub API).
(Market) Return (t)	Returns from date t (CRSP).
Return-on-Assets	Net income divided by lagged total assets (Compustat).
Sales Growth	Annual percentage change in sales (Compustat).
Scope	Number of industries in which the firm operates, see Hoberg and Phillips (2025) (Hoberg-Phillips Data Library).
Stars	Number of stars of a repository as of February 2024 (GitHub API).
Tangibility	Property, plant, and equipment scaled by total assets (Compustat).
Template	Indicator variable that equals one if the repository is configured as a template, which allows copies to be created without retaining the commit history (GitHub API).
Trademark First Use	Indicator value that equals one if a trademark is first in commercial use on that date (USPTO).
(Idiosyncratic) Volatility	Standard deviation of daily returns over one month. Idiosyncratic volatility is similarly defined using returns net of market returns (CRSP).
ξ	An estimate of the private value of repositories (in 2023 dollars) using stock market returns around the repository release date.
ξ^{alt}	An alternative estimate of the private value of repositories (in 2023 dollars) using stock market returns around the repository release date that does not make assumptions about the distribution of repository values.