The Private Value of Open-Source Innovation*

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

August 2024

Abstract

We investigate open-source innovation by public firms and the private value it generates for these firms. Unlike patents, which grant inventors exclusive rights to their inventions, open-source innovations can be used by anyone. Nevertheless, using an extensive dataset of public-firm activity on GitHub, we find that open-source activity is widespread. Firms with open-source projects represent 68% of the U.S. stock market across 86% of industries. We estimate the value of all projects in our sample to be nearly $25 billion, with the average project generating $832,000. Firms facing less competition generate more value, suggesting they capture a larger portion of the total value of their innovation. Open-source value significantly predicts firm growth, but it also stimulates patenting by other firms and does not result in creative destruction. These results contribute to our understanding of how innovation generates private value in the absence of excludability.

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

*We thank Fabrizio Corre, Mikael Paaso, and seminar participants at Erasmus University Rotterdam for helpful comments. We are responsible for any remaining errors. Emails: emery@rsm.nl, clim26@buffalo.edu, ye@rsm.nl.

†Rotterdam School of Management, Erasmus University Rotterdam.

‡School of Management, University at Buffalo.

1 Introduction

Innovation is costly to produce but can be used by multiple parties with little or no additional cost. Consequently, inventors bear the full cost of innovation but receive only a portion of the benefits. In order to incentivize investment in innovation, systems (e.g., patents) have been created to grant inventors exclusive rights to monetize their innovation for a given period of time. This excludability is seen as crucial for deriving private value from innovation (Arrow, 1962; Crouzet et al., 2022). However, the past decade has seen a rise in so-called open-source innovation. When an innovation is "open-sourced," it is made publicly available to all parties at little or no cost. Perhaps surprisingly, many firms choose to make their innovation open source. A recent survey finds that 90% of Fortune 100 companies use GitHub, the largest platform for developing open-source innovation.[1] However, it remains unclear what private value these profit-maximizing entities derive from making their costly innovation freely available.

In this paper, we study the open-source innovation of publicly traded firms. 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 tends to choose to develop their innovation via open source. Second, we estimate the private value of open-source innovation and investigate the innovation, firm, and product market characteristics that most strongly correlate with private value. Finally, we examine the relation between open-source innovation and future firm growth, as well as the interplay between the creative destruction inherent in the innovation process and the positive externalities created by open-source innovation.

There are many ways in which open-source innovation may create private value for firms. We group these ways into three main mechanisms.[2] First, making innovation free to use can

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

[2] 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), Dahlander and Gann (2010), Henkel et al. (2014), Parker et al. (2017), Alexy et al. (2018), Nagle (2018), Teece (2018) and Lin and Maruping (2022). For reviews of the open-source literature, see von Hippel and von Krogh

maximize the adoption of that innovation. Increased adoption can create value by improving innovation through community development, giving the firm more control over subsequent development of the technology, creating a product ecosystem that deters customers from switching to competitors, and increasing demand for complementary proprietary products. These phenomena constitute network effects whereby the value of the innovation increases exponentially with the number of users. Second, open-source innovation may provide value through labor considerations. Valuable employees may want to share their work to increase their reputation within the open-source community, and firms may identify talented workers who contribute to open-source projects (and have lower integration costs due to already being familiar with the firm's infrastructure). Finally, making innovation open source may enhance the firm's reputation. Firms that contribute to open-source projects may be seen as more community-oriented and open-source projects may be seen as more transparent and certified for quality.

We study open-source activity using 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 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 produce open-source innovation on GitHub, those firms represent 68% of the total stock market capitalization and 80% of the total research and development (R&D) expenditure by public firms in 2023. In comparison, such firms represent only 20% of the total stock market capitalization at the beginning of our sample. Moreover, while 32% of firms producing open-source innovation on GitHub are from the "Computer Software" industry, 86% of industries have at least one such firm,[3] demonstrating the growing scope of open-source innovation. We find that firms that contribute to open-source projects are larger, more valuable, more innovative, and face less competition on average. However, in

(2003), Goldfarb and Tucker (2019), and Dahlander et al. (2021).

[3] This figure is based on the Fama-French 49 industries classification system.

a regression setting, most of these differences are absorbed by firm fixed effects, suggesting that firm fixed effects can account for much of the potential selection bias in the choice to make innovation open source.

We next employ a modified version of the method developed by Kogan et al. (2017) to measure the private value of repositories as estimated by investors. The methodology is based on observing firm-specific stock returns over the three days following the release of a repository. Using this methodology, we find that investors estimate the average repository in our sample to be worth $832,260 (in 2023 dollars), with average values increasing significantly over the sample period. The total private value created by repositories in our sample is nearly $25 billion. The most valuable GitHub portfolios are owned by Amazon.com, Inc and Microsoft Corp, both valued at nearly $8 billion, and repositories using Python as the main programming language produce the most total value in our sample.

Attributing stock returns around the repository announcement to the repository assumes that investors respond to the release of repositories on GitHub. Otherwise, the estimates of repository value only reflect noise in the market. To assess this possibility, we regress repository value on a measure of future repository popularity. We find that more valuable repositories end up being significantly more popular in the future, which suggests that stock price reactions to repository announcements contain significant information. To provide further validation, we perform a placebo test where we assign random announcement days, within the true announcement year, to each repository and estimate a placebo value. We then perform the same regression of (placebo) repository value on future repository popularity. In none of the 500 iterations of this tests did the placebo relation reach the economic and statistical levels of the true relation.

Investigating the determinants of open-source value, we find that repositories with copyleft licenses, which place restrictions on commercial use of the repository, are slightly more valuable than repositories with fully permissive licenses. However, the difference is not statistically significant. It therefore appears that both the ability to exclude competitive use and promote further adoption are valuable for open-source projects and represent an important trade-off for firms. We also find that larger repositories (e.g., more lines of code) are not necessarily more valuable and repositories with more subsequent issues opened (e.g., bugs) are perceived as less valuable when initially released.

Since a firm's competitors are most likely to benefit from the open-source nature of GitHub repositories, 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 have repositories that produce more private value. This is likely a function of these firms being able to capture a larger portion of the total value created by the repository, which is the sum of the private and public value. We also find that, controlling for the level of competition, firms 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.

Finally, we investigate the relation between open-source innovation and future firm growth. We find that more-valuable repositories are associated with a larger growth in sales, profits, market share, number of employees, and both the number and value of patents granted over the following three years. Thus, open-source innovation produces significant value for the innovator despite it being available for use by competitors. Furthermore, these results demonstrate a complementarity between open-source innovation and patenting, which is consistent with the increased adoption driven by open-source innovation increasing the profitability of patentable innovation.

We also investigate the relation between competitor open-source innovation and future firm growth. Many theories of innovation highlight the creative destruction that occurs when one innovation makes another innovation obsolete (Schumpeter, 1912; Aghion and Howitt, 1992). Consistent with these theories, Kogan et al. (2017) find that competitor patenting activity is negatively associated with future firm growth. In the context of open-source innovation, however, competitors can also use the innovation and so share in its value. These positive externalities create an ambiguous relation between competitor open-source innovation and future firm growth. Thus, in contrast to the results for patents documented in Kogan et al. (2017), we find no significant relation between competitor open-source innovation and future firm growth. This result is consistent with the negative externalities of creative destruction and positive externalities of open source negating each other. We do, however, observe an increase in the number of patents granted to the firm over the following three years. This result provides suggestive evidence that open-source innovation can potentially promote follow-up innovation.

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 typically explored the value of innovation within traditional intellectual property protection 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., 2023; Ahmadi et al., 2024). We explore how innovative outputs contribute to a company's value even when freely disclosed to a broad audience through open-source licenses. Specifically, in the case of open-source software, as noted by Lerner and Tirole (2005a), the contribution of intellectual assets created by a company to its value creation and future growth is often indirect, making it challenging to measure quantitatively. We address this challenge by leveraging financial markets to measure the value of intellectual property without excludability through the value of repositories in open-source platforms.

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. Furthermore, it allows us to develop a new stock-market-based measure of the value of open-source innovation. Previous research has estimated the value of particular open-source software, such as Apache and nginx, using the cost of replicating 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 the length of software code and labor costs (Robbins et al., 2021; Blind et al., 2021), which may not directly reflect its contribution to firm value. For example, Hoffmann et al. (2024) use this strategy to estimate the cost of replicating the most-used open-source software either once ($4.15 billion) or individually by all firms that use it ($8.8 trillion). 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 at the firm level and explore heterogeneity therein.

We also contribute to the literature on innovation and firm growth by showing that the value of open-source innovation provides significant insights into firm growth beyond what is captured by the measure of patent value proposed by Kogan et al. (2017). Previous research has shown that companies can enhance their software development capabilities, firm productivity, or access to venture capital by being active in 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 a company's long-term growth in product sales and innovative output. Furthermore, we examine the relation between a firm's own open-source innovation, its competitors' open-source innovation, and future firm growth.

More broadly, we contribute to the literature that examines the optimal intellectual property protection system to promote innovation in the overall economy. Hoffmann et al. (2024) and Bessen and Maskin (2009) argue that the cumulative nature of innovation is crucial to how research should be rewarded. Specifically, Bessen and Maskin (2009) states that the more cumulative and sequential the innovation, the less useful patent protection becomes for encouraging innovation, and society, as well as inventors themselves, may be better off without such protection, with the software industry highlighted as an example. Galasso and Schankerman (2015) provides empirical evidence that the positive effect of patent invalidation on follow-on innovation is more substantial when innovation is sequential, such as in the computers and electronics industries. By documenting the private value of open-source innovation, our paper demonstrates that when innovation is highly cumulative and sequential, open-source licenses can serve as an alternative system that generates company value and growth while not hindering downstream innovation.

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.

2.1 Initiate open source projects

To deploy their projects on GitHub, firms need to create organization accounts. While some firms create only one organization account, others create multiple organization 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. Notably, even though there had been additional costs for adding private repositories before GitHub pricing model was changed from repository-based to user-based in 2015, GitHub has provided free hosting for public repositories since its inception.

One crucial 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.[4] The choice of license can have different implications for commercial use. There are two primary categories of open-source licenses based on their permissiveness: permissive licenses

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

and copyleft licenses. Permissive licenses impose minimal restrictions on how the source code can be used. In contrast, copyleft licenses 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. It is worth noting that 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.[5] Internet Appendix A compares different types of licenses based on permissions and conditions.

2.2 Project development and community interaction

GitHub operates on the Git system that provides a collaborative and distributed approach to software development. In this context, several key processes and community interactions play a pivotal role in driving innovation and progress.

The development process begins with developers making changes to the codebase. These changes are committed locally, along with brief summaries explaining the nature of the modifications. Subsequently, these developers push these commits to remote branches, ensuring that other contributors and users can view the updates.

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 the changes made in this personal fork are deemed valuable and applicable to the original

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

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 on GitHub activities of U.S. public firms, we begin by linking GitHub organization accounts with firms. Following the methodology of Conti et al. (2021), 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 168,085 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, pushes, and pull requests. Our panel spans between 2015 and 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 140,824 repositories extant as of February 2024. This includes an array of attributes from descriptive repository metadata, such as creation dates, licenses, and programming languages, to quantitative measures of community engagement, including the number of stars, watchers, and forks.

3.2 Summary statistics

Before moving to the estimation of open-source value, we first provide an overview of open-source activity. We begin by documenting trends in open-source engagement during our sample period, which are plotted in Figure 1. The dashed yellow line plots the cumulative number of repositories created by firms that are public as of that month, which totals 122,107 repositories by the end of our sample period.[6] The blue line represents the percentage of public firms that have created at least one repository on GitHub as of that month (henceforth "open-source firms"). This percentage increases steadily from 4.8% in January 2015 to 18.1% in December 2023. The red line plots the percentage of total market capitalization represented by open-source firms, which is 67.5% at the end of our sample period. Finally, the green line plots the percentage of total R&D expenditure represented by open-source firms, which is 80.2% at the end of our sample period. Thus, despite open-source firms being only one-fifth of all public firms, they represent two-thirds of the stock market and

[6] Note that the cumulative counts of repositories in our firm-month panel are slightly smaller than the original matched sample for two reasons. First, some publicly visible repositories that never appear in major open-source event records (such as issues, pushes, and pull requests) are excluded from the panel. Second, we exclude delisted firms starting from the month of delisting.

over four-fifths of investment in innovation. We therefore conclude that firms engaged in open-source activities are an important part of the US economy.

Next, we examine the distribution of open-source firms across industries. Figure 2 presents pie charts of this distribution at the firm and repository levels. We use the Fama-French five industry classification scheme (Fama and French, 1997) and further separate out the "Computer Software" and "Finance" industries as defined by the Fama-French 49 industry classification scheme. One may assume that software firms represent the majority of open-source firms in our sample. However, we find that only 32.2% of open-source firms in our sample come from the "Computer Software" industry. This said, more than two-thirds of repositories in our sample are owned by firms in this industry, confirming the intuition that software firms are the most active in open-source innovation. Nonetheless, other industries are also reasonably well represented in our sample, particularly the "Business Equipment, Telephone and Television Transmission" and "Consumer Durables, NonDurables, Wholesale, Retail, and Some Services" industries.

We provide further summary statistics on 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. As can be inferred from Figure 2, open-source activity is most common among "Computer Software" firms, with 62.6% of firms in this industry having open-source activity. The "Healthcare, Medical Equipment, and Drugs" industry is the least active, with only 6.4% of firms having open-source activity. This result may be a reflection of the importance of excludability for innovation in this industry.[7]

Panel B compares firm and product market characteristics for firms with and without open-source activity. The panel reports the mean and median values of each group of firms over our sample period. Firm characteristics include the number of employees, market-to-book ratio, return-on-assets, investment, sales growth, tangibility, and research and development (R&D) expenditure scaled by total assets, all of which are calculated using data from

[7] E.g., see https://www.reuters.com/legal/litigation/us-senators-ask-regulators-clear-drug-patent-thickets-2022-06-08/.

Compustat. We also calculate market capitalization and annual returns using data from the Center for Research in Security Prices (CRSP) and obtain data on patent portfolios from Kogan et al. (2017).

We examine the product market characteristics of open-source firms because technological spillovers to competitors represent the largest potential negative externality faced by open-source firms. Product market characteristics include market power, scope, product market centrality, product market similarity, and product market fluidity. Market power measures a firm's dominance within its product market, which we proxy for using a structural estimate of markups from Pellegrino (2024). Scope measures the number of industries in which the firm operates, based on their product descriptions in SEC filings (Hoberg and Phillips, 2024). Product market centrality is calculated as the eigenvector centrality of a firm in the product-market network, which is constructed using similarity scores from Hoberg and Phillips (2016). This quantity reflects the extent of competition faced by the firm, but it also measures the extent to which the firm benefits from spillover effects in the network (i.e., network effects), which can be crucial for the success 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 than non-open-source firms on average, based on market capitalization, employees, and number of patents. These firms also tend to have higher valuations, based on market-to-book ratio, which could reflect investors' assessment of growth opportunities resulting from the firms' innovation. Intuitively, open-source firms tend to have less tangible assets and larger R&D expenditures. Finally, open-source firms appear to 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 do not account for the concentration of open-source activity in certain industries, particularly the "Computer Software" industry. The observed differences between open-source and non-open-source firms could, therefore, be a function of industry differences rather than firm-specific characteristics. We investigate this possibility in the next section.

3.3 Determinants of open-source activity

To provide a more rigorous characterization of open-source firms, we next consider 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 key 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 2. 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.[8] The results in Column (1) therefore reflect the average difference between open-source and non-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

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

peers (Return-on-Assets) and have lower annual returns. One possible interpretation is that open-source firms are less profitable because they have greater investment in R&D (and in open-source innovation specifically).

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.

4 Open-Source Value

Having characterized open-source firms and assessed the determinants of open-source activity, we next turn to estimating the economic value of open-source innovation. Specifically, we leverage financial markets to estimate the value, in dollars, of GitHub repositories based on stock returns around the date on which the repository was made public. In the following, we outline the procedure to estimate repository 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 value

Our procedure for estimating the value of repositories closely follows the procedure developed by Kogan et al. (2017) to estimate patent value and used by Desai et al. (2023) to estimate trademark value. We provide a detailed discussion of this procedure in Internet Appendix B and briefly outline the crucial points in this section.

The procedure involves observing stock returns in the three-day window following the announcement of the repository, $[t,\ t+2]$. We cumulate market-adjusted returns over the three-day announcement window for repository i, which we label R_i. We assume that R_i is a function of both investor reaction to the repository announcement, v_i, and idiosyncratic noise.

We construct the estimate of repository value as the product of the investor reaction to the repository announcement and the firm's market capitalization on the day prior to the announcement. 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

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

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 announcement 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 announcement.

Internet Appendix B 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. 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.

4.2 Summary statistics

We estimate Equation (1) for the 30,034 original repositories that have available announcement dates as well as the required stock return data from CRSP.[9] In Panel A of Table 3, we report the mean, standard deviation, and multiple distribution percentiles (1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th) of several variables. The mean (median) three-day cumulative market-adjusted announcement return (R_i) is 0.20% (0.08%) and the mean (median) expected return attributable to the repository announcement ($E[v_i|R_i]$) is 0.27% (0.13%). The difference between the mean and median for both variables indicates that market reactions to repository announcements are positively skewed.

We find that the mean value, ξ, for repositories in our sample is \$832,260, and the median value is \$548,444. There is also significant skewness in this variable, with the 99th percentile of repository value exceeding \$5,000,000 and the most valuable repositories exceeding \$12,000,000 (untabulated). These values are reported in 2023 dollars. For comparison, the mean value for patents granted over a similar period (2015-2022) is \$48 million in 2023

[9] We focus on original (i.e., not forked) repositories because the release date for forked repositories is less clearly defined.

dollars, as calculated using data from Kogan et al. (2017). Given that patented innovation typically requires higher investment and benefits from the excludability provided by the patent system, we consider the relatively lower mean repository value to be plausible. We also report summary statistics for the value of fully permissive repositories, which represent 69.6% of our sample. These repositories tend to be slightly more valuable, with a mean value of \$880,752 and a median value of \$643,956. Finally, 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, with the mean repository being starred 209 times but the median repository being starred only nine times.

We also report the distribution of repository values for different industries in Panel B of Table 3. The majority of the repositories in our sample are created by firms in the "Computer Software" industry (57.4%) and the average value for repositories in this industry (\$784,464) is similar to that of the complete sample. On average, repositories from the "Consumer Durables, NonDurables, Wholesale, Retail, and Some Services" industry are the most valuable (\$1,093,772) and make up the second largest group of repositories (24.2%).

Panel C of Table 3 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, both nearing \$8 billion. These two portfolios alone represent 62.5% of all repository value and 45.6% of all repositories in our sample. For both firms, the majority of their repositories have fully permissive licenses. 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, we find that the repositories in our sample generated nearly \$25 billion of private value for public firms.

Panel D of Table 3 reports the 10 programming languages that generate the most value as classified by each repository's main programming language. Python is the most commonly represented language (22.3% of repositories) and is associated with the most total repository value (\$6,876,648,682). Python also has the highest average and median repository value among the languages listed.[10,11] Go and JavaScript are notable for having the largest skewness in repository value among the languages listed, and C++ and HTML are notable for having the smallest percentage of repositories with a fully permissive license among the languages listed.

We also investigate how the average repository value as evolved over time. In Figure 3, we plot the average ξ, in 2023 dollars, of repositories released each quarter from 2015 through 2023. Average repository values hovered around $400,000 from 2015 through 2017 and then jumped to between $600,000 and $800,000 from 2018 through 2019. This increase coincides with Microsoft's acquisition of GitHub, which was announced in June of 2018. Average repository values significantly increased again the first quarter of 2020, which is likely attributable to expectations of increased digitalization resulting from the COVID-19 shutdown. Since 2020, average repository 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 characteristics most strongly correlate with open-source value. We investigate a broad set of repository, firm, and product market characteristics to give an extensive assessment of these correlations. 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 C reports univariate correlations among all pairs of variables included in our analysis. We view our results in this section as descriptive in nature and intended to provide a more complete characterization of open-source value.

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

[11] Other languages with a higher average repository value and at least 10 repositories include Cuda ($2,335,779, 23 repositories), Bicep ($1,472,982, 79 repositories), Swift ($1,441,120, 359 repositories), and CMake ($1,331,631, 25 repositories).

4.3.1 Repository popularity

We begin by investigating repository popularity. While we contend that the statistics reported in Table 3 indicate our measure of repository value is reasonable, there is still the possibility that significant noise in the estimation procedure renders the estimates largely uninformative. If this is the case, then we would expect repository value to be at most weakly correlated with subsequent repository popularity. We therefore investigate this possibility to provide validity for the estimates of repository value.

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 external user, which allows the user to stay updated on any changes made to the repository and indicates significant interest in the repository. We regress the natural logarithm of repository value, ξ, on the natural logarithm of one plus the number of stars the repository has received.[12] We control for the natural logarithms of market capitalization (measured as of the repository announcement), volatility (measured over the announcement year), employees, and patent-portfolio value (both measured as of the prior year). We also include various fixed effects depending on the specification, including year, industry (at the three-digit SIC level), industry-year, firm, and firm-year fixed effects. We double-cluster standard errors by year and industry and all independent variables are standardized to facilitate interpretation.

The results of these regressions are reported in Table 4. In Column (1), we report the regression with year fixed effects and find a significant correlation between repository value and subsequent popularity. The regression reported in Column (2) adds industry fixed effects and the regression reported in Column (3) replaces these fixed effects with industry-year fixed effects. In both cases, we continue to find a significant correlation between repository value and subsequent popularity, with the correlation generally increasing in significance as the fixed effects become stricter. Economically, the estimate reported in Column (3)

[12] We take the natural logarithm of each variable to adjust for the skewness documented in the previous section.

indicates that repositories that end up being one standard deviation more popular have a 9.8% higher valuation when released. Finally, Column (4) reports the regression with firm fixed effects, and Column (5) includes firm-year fixed effects. Both results demonstrate a strong correlation between within-firm repository value and subsequent popularity.

We therefore conclude that repositories that are estimated to be more valuable when they are released tend to be significantly more popular in the future. This result indicates that the estimation procedure produces informative estimates of open-source value.

4.3.2 Repository, firm, and product market characteristics

We next investigate the correlations between repository value and other repository characteristics, firm characteristics, and product market characteristics. The results are reported in Table 5 with Panels A, B, and C corresponding to each category of characteristics, respectively. Each regression includes 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 reporting the regression including all characteristics from that category.

In Panel A, we investigate repository characteristics. We first consider the type of license covering the repository. Each repository is classified into one of three groups based on its license: permissive (no restrictions on use), copyleft (some restrictions on use), and other (cannot be classified). We then include indicator variables for "copyleft" repositories and "other" repositories in the regression such that "permissive" repositories represent the omitted category. Traditional models of the economics of innovation suggest that excludability increases the private value of innovation (Schumpeter, 1912), and Lerner and Tirole (2005b) explicitly discuss how license restrictiveness increases open-source value. Consistent with these theories, we find that copyleft repositories are approximately 8.7% more valuable, on average, than permissive repositories (Column (6)), however this difference is not statistically significant. Thus, there appears to also be value in allowing unrestricted use of the repository, which is consistent with increased adoption being an important driver of open-source value.

We next examine repositories designated as templates. Template repositories can be easily duplicated into new repositories with identical directory structure, branches, and files without keeping the commit history. Given that these repositories are less likely to contain a unique innovation, we expect them to be less valuable on average. Consistent with this notion, we find that template repositories are approximately 25.9% less valuable, on average, than non-template repositories (Column (6)).

We also examine the byte size of repositories. Repositories with more lines of code will have a larger byte size, all else equal, and may therefore be more valuable. However, we find a negative relation between repository size and repository value that is marginally statistically significant when included by itself (Column (3)) and statistically insignificant when controlling for other repository characteristics (Column (6)). This result could be due to other files included in the repository, such as images, that increase the repository size and do not represent additional lines of code. To investigate this possibility, we examine the subset of repositories (2,223 repositories) for which byte size is separately categorized into binary (e.g., images) and non-binary (e.g., lines of code) data. However, we still find that both types of data are negatively and insignificantly related to repository value (untabulated).[13] We therefore conclude that larger repositories are not necessarily more valuable.

We then examine the number of repositories previously released by the firm. This quantity is negatively and significantly related to repository value, suggesting that firms producing a lower quantity of repositories tend to produce higher-quality repositories. Finally, we examine the cumulative number of issues opened for the repository as of December 31, 2023. While this value captures potential bugs or problems with the repository, it also scales with the overall popularity of the repository. However, we control for repository popularity in the

[13] We do, however, find that the ratio of non-binary byte size to total byte size is positively and significantly related to repository 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).

regression using the number of stars, and thus interpret issues opened as a negative reflection of repository quality. Consistent with this interpretation, we find a negative and significant relation between the number of issues opened in the future and repository value estimated by investors at announcement.

In Panel B of Table 5, 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 the majority of these variables are not significantly related to repository value in the cross-section. The two exceptions are return-on-assets, which suggests that more profitable firms tend to produce more valuable repositories, and the indicator variable for missing R&D expenditure, which suggests that such firms also tend to produce more valuable repositories. The latter result could imply that investment in open-source innovation is difficult for firms to quantify as R&D expenditure. At the least, the result demonstrates that R&D expenditure does not fully capture open-source activities.

Finally, in Panel C of Table 5, we investigate 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, as measured by the estimate of markups developed by Pellegrino (2024). Market power is negatively related to competition in theory and reflects a firm's ability to extract rents from its customers. Firms with more market power may face less risk of their repositories being used by competitors and may be able to extract more value from the increased adoption resulting from the repository being free to use. Consistent with these notions, we find that firms with more market power tend to produce more valuable repositories.

We then consider the firm's centrality in the product market network. To the extent that this quantity reflects the level of competition faced by the firm, similar to market power, we would expect centrality to be negatively related to repository value. However, centrality also measures the extent to which the firm benefits from network effects in the product market, which we expect to be positively related to repository value. Consistent with these opposing predictions, we find that centrality is insignificantly related to repository value when included by itself in Column (2). However, this relation becomes positive and significant in Column (6) when including other product market characteristics that also reflect competition (e.g., market power). The result in Column (6) therefore supports the hypothesis that network effects are a significant driver of value for open-source innovation.

Finally, we also consider the scope, product market similarity, and product market fluidity of firms. We find that scope is negatively related to repository value, suggesting that firms with a sharper product focus tend to produce more valuable repositories. We also find that product market fluidity is negatively related to repository value, suggesting that repositories create more valuable when the firm is in a more stable product market.

In summary, the results for product market characteristics suggest that repository value is negatively related to competition. However, it is important to note that our estimates reflect private value. It is therefore possible that firms facing less competition are able to capture a larger fraction of the total value created by the repository, which is the sum of private and public value.

It is also important to note that repository popularity is positively and significantly correlated with repository value across all regressions reported in Table 5. This result further supports the validation exercise from the previous section. To mitigate any further concerns of spurious correlation between repository popularity and repository value, the next section performs a placebo test on the announcement date of repositories.

4.4 Placebo test

To provide further validation for the estimates of repository value, we perform a placebo test on the repository release date. One may question whether investors pay attention to the releases of repositories on GitHub, and if they do, whether these repositories significantly affect investors' demand such that it produces a measurable market reaction. Alternatively stated, are the market reactions we observe, and the correlation with subsequent repository popularity, simply spurious?

We investigate this possibility by randomly assigning each repository a placebo release date in the same year as the true release date. We then estimate the placebo value of the repository based on the market reaction on this placebo release date. Finally, we regress repository placebo value on the true number of stars subsequently received by the repository. The regression specification follows that of the previous section with additional control variables included based on their statistical significance in Table 5.[14] We repeat this process 500 times. The resulting distribution of coefficient estimates, corresponding to number of stars, and t-statistics are plotted in Panel A and Panel B of Figure 4, respectively. Each panel also plots a vertical dotted line corresponding to the results from an identical regression with the true repository values.[15]

It is readily apparent in the figure that the true coefficient and t-statistic are significant outliers relative to the placebo estimates. We therefore conclude that the market reactions on repository release days are in fact, on average, directly related to the repository release.[16]

5 Open Source Innovation and Firm Growth

Technological innovations are recognized for driving long-term growth for companies, while innovations by other firms are often associated with a negative impact, known as

[14] Specifically, the baseline specification controls for stock market capitalization, stock volatility, employees, and total patent value, includes industry-year fixed effects, and double-clusters standard errors on industry and year. The additional control variables include indicator variables for license type, an indicator variable for whether the repository is a template, the number of repositories previously released by the firm, the cumulative number of issues opened for the repository as of December 31, 2023, return-on-assets, R&D expenditure scaled by total assets and an indicator variable that equals one if R&D expenditure is missing, market power, scope, product market centrality, and product market fluidity.

[15] In the true regression, the coefficient on number of stars is 0.061 with a t-statistic of 3.91.

[16] While there are undoubtedly minor repositories that do not illicit a measurable market reaction, these repositories are by-definition less valuable, and this fact is reflected in the estimated value.

creative destruction (e.g., Kogan et al. (2017)). However, this conclusion in the literature is grounded in the traditional approach to protecting intellectual property, namely granting exclusive rights to use the technology for a specific period. In contrast, open-source licenses grant usage rights to the general public, including a firm's competitors, complicating the prediction of the relation between firm growth, open-source innovation, and creative destruction.

To investigate this question, we study the relation between a firm's future growth and the open-source innovation of the firm and its competitors. In particular, we calculate the firm-level repository value, $\xi_{f,t}$, as the sum of the repository values ξ_i for all repositories posted by firm f in year t. Similarly, we calculate the aggregate value of repositories posted by a firm f's competitors as the sum of repository values at the industry level (defined at the SIC3 level) in year t, excluding the repositories posted by firm f during the same period. We consider several dependent variables (Y) including the growth of product sales, profits, market share, number of employees, and number and value of patents:

$$\ln Y_{f,t+k} - \ln Y_{f,t} = \beta_k \ln(\xi_{f,t} + 1) + \gamma_k \ln\left(\sum_{f' \in I \setminus f} \xi_{f',t} + 1\right) + \psi_k X_{f,t} + \epsilon_{f,t+k}, \qquad (2)$$

where the horizon k varies from one to three years. The vector X includes the natural logarithms of market capitalization, volatility, number of employees, patent-portfolio value, and $Y_{f,t}$. We also include industry-year fixed effects (at the three-digit SIC level).

The results are outlined in Table 6. In this analysis, our focus is on the coefficients β_k and γ_k, which capture the impact of innovations by the firm and its competitors (creative destruction) on firm growth. The initial three columns present our estimates of β_k for $k = 1, 2, 3$. We observe a significantly positive relation between a company's open-source innovation and its future growth across all time horizons. In unreported results, we confirm that open-source innovation does not affect the growth of the company's tangible assets. Overall, our findings suggest that, even when the repositories with open licenses are accessible to others, firms still experience benefits from such innovation.

By contrast, the coefficients γ_k are largely insignificant for all dependent variables. This stands in stark contrast to the notable creative destruction associated with innovation measured by patents reported in Kogan et al. (2017). While competitors' open-source innovation may generate both positive and negative externalities for a firm, our results indicate that these externalities may offset each other, leading to an insignificant coefficient. The one exception is the number of patents granted up to three years in the future, which is positive and statistically significant. This result suggests that competitors' open-source innovation may stimulate patentable innovation for the firm, which is consistent with the open-source nature of this type of innovation facilitating further innovation.

6 Conclusion

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

We find that open-source engagement is highly prevalent in the U.S. economy. Despite only 18% of public firms having open-source projects, those firms represent 68% of stock-market value and 80% of R&D expenditure across 86% of Fama-French 49 industries. Firms with open-source projects tend to be larger, more valuable, more innovative, and face less competition on average.

We estimate the value of open-source projects based on stock-market reactions. The average project in our sample is valued at $832,260 (in 2023 dollars), with average values exceeding $1,300,000 by the end of 2023. The total private value created by all projects in our sample is $25 billion. We find that firms face a trade-off in the values of restricting commercial use and maximizing adoption of their projects. Furthermore, larger projects (e.g., more lines of code) do not necessarily create more private value and firms facing less competition appear able to capture a larger portion of the projects total value as private value.

Finally, we find that valuable open-source innovation predicts future firm growth in terms of sales, profits, market share, number of employees, and both the number and value of patents. We do not find evidence of open-source innovation resulting in creative destruction, potentially due to the positive externalities of the open-source model offsetting the negative externalities of depreciating innovation. We do, however, find that open-source innovation stimulates patenting by competitors, consistent with open-source promoting follow-up innovation.

In summary, these results provide new evidence on the value of innovation without excludability. Our estimates of open-source value open up avenues for future research on the benefits firms gain from open-source innovation and the valuation of intangible assets. More broadly, our results imply that open-source licenses can complement conventional intellectual property protection systems for certain types of intellectual property, thereby enhancing firm value and improving societal welfare. Future research exploring which industries and types of goods benefit most from this combined system could yield important policy implications.

References

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., 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.

Bessen, J. and E. Maskin (2009). Sequential innovation, patents, and imitation. *RAND Journal of Economics 40*(4), 611–635.

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

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

Conti, A., C. Peukert, and M. Roche (2021). Beefing IT up for your Investor? Open Sourcing and Startup Funding: Evidence from GitHub. Working paper, IE Business School, Harvard University, University of Lausanne.

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). How open is innovation? *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 (2023). 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.

Galasso, A. and M. Schankerman (2015). Patents and cumulative innovation: Causal evidence from the courts. *Quarterly Journal of Economics 130*(1), 317–369.

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

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

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

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.

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 (2024). Scope, Scale and Concentration: The 21st Century Firm. *Journal of Finance*. Forthcoming.

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.

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

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.

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.

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. (2024). Product Differentiation and Oligopoly: a Network Approach. *American Economic Review*. Forthcoming.

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.

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.

von Hippel, E. and G. von Krogh (2003). Open Source Software and the "Private-Collective"' Innovation Model: Issues for Organization Science. *Organization Science 14*(2), 107–225.

Figure 1
Trends in Open-Source Engagement Among U.S. Public Firms (2015-2023)

This graph plots the time series U.S. firms' participation in open-source activities through the creation of public GitHub repositories from 2015 to 2023. It represents the proportion of firms making repositories public in terms of total number of firms, market capitalization, and R&D expenditure (left y-axis), and it tracks the cumulative number of public repositories owned by these firms (right y-axis).

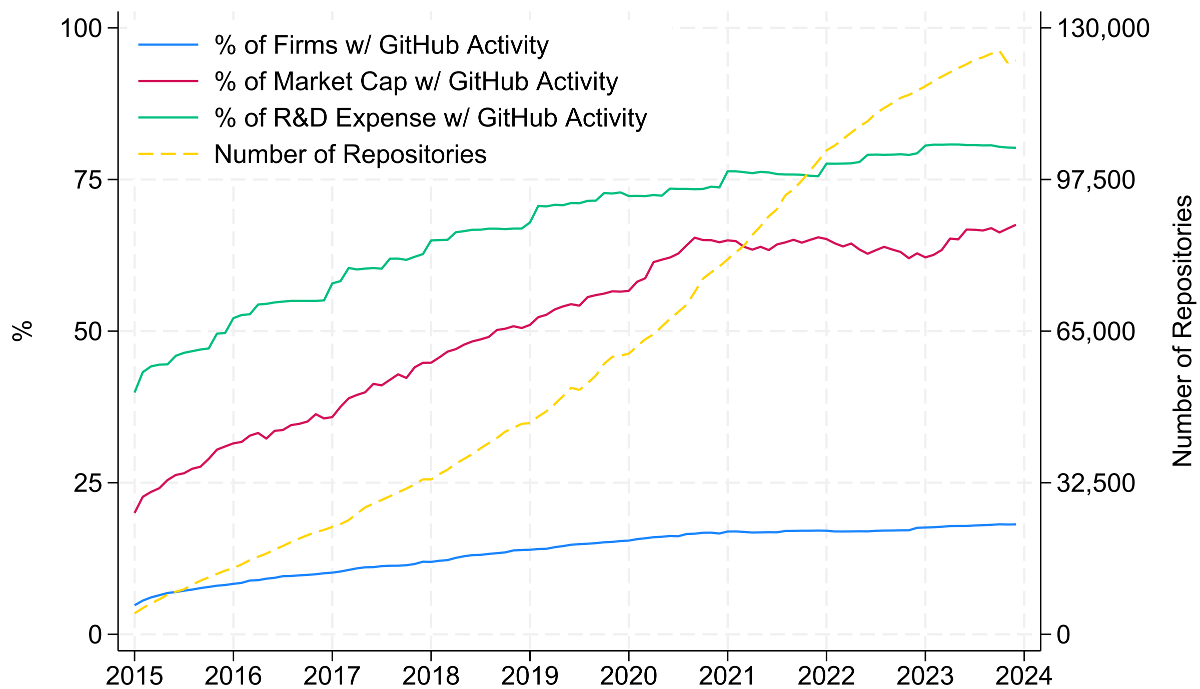

Figure 2
Industrial Distribution of Open-Source Engagement

This figure illustrates the percentage of firms with GitHub activity and their respective repositories across various industries over the period from 2015 to 2023. The percentages are derived by dividing the number of firms with GitHub activity in a given industry by the total number of firms active on GitHub. The left pie chart shows the proportion of firms with any GitHub activity, categorized by industry, while the right pie chart displays the distribution of all repositories owned by these firms. Industry classification adheres to the Fama French 5 Industries, with the exception that computer software is distinctly separated from the business equipment, telephone and television transmission industry.

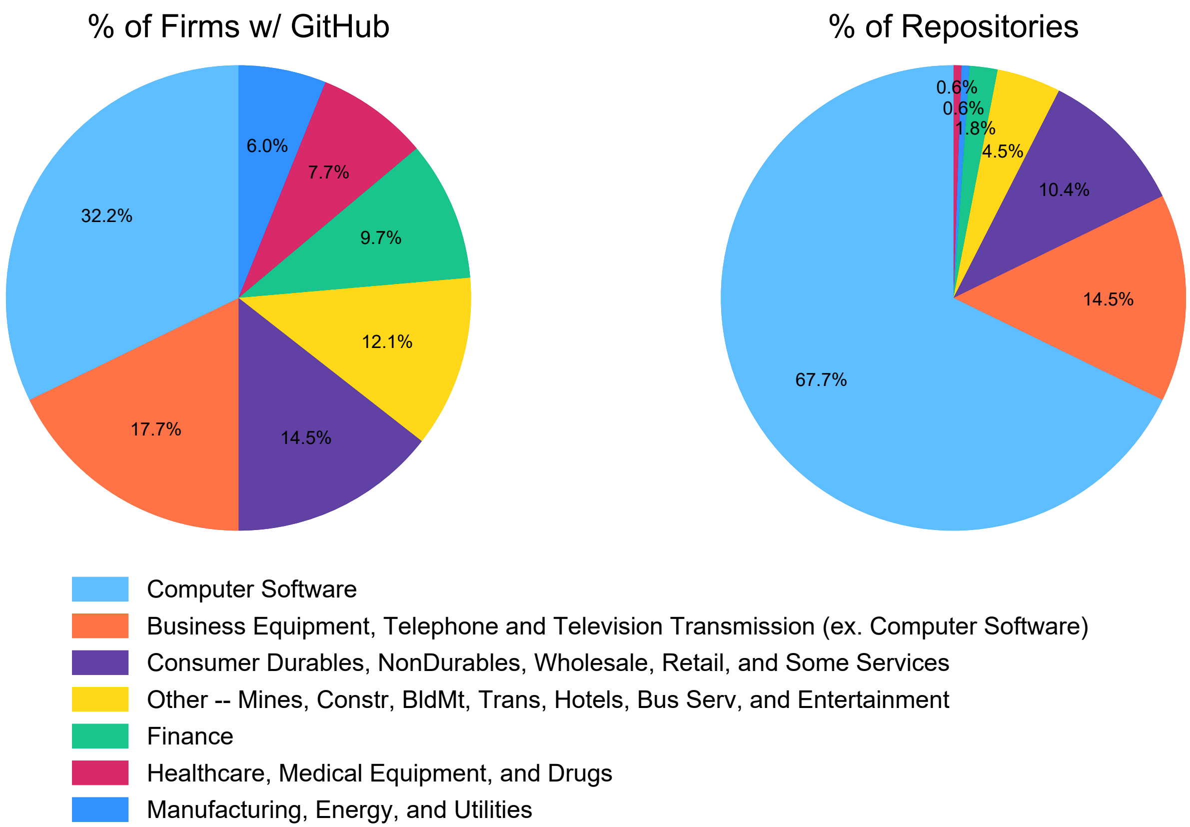

Figure 3
Average Estimated Repository Value by Quarter

This figure displays the trajectory of the average repository value, ξ, from 2015 to 2023. The values are computed using the methodology detailed in Section 4.1 and have been adjusted to reflect 2023 dollar values.

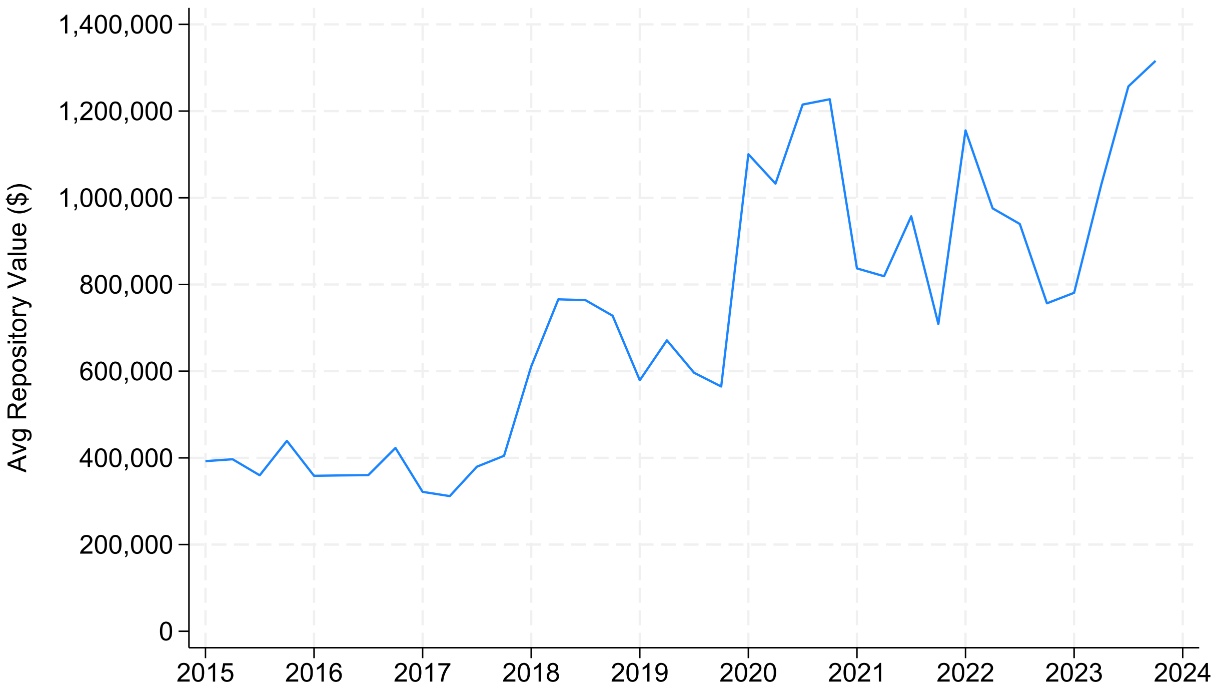

Figure 4
Placebo Test Results

This figure displays the distributions of coefficient estimates and t-statistics from 500 iterations of placebo tests, conducted to validate the value measure for repositories. In the procedure, each repository is assigned a random placebo release date within its actual release year, and its placebo value is determined by the market response on that date. Panel A shows the distribution of coefficient estimates linked to the number of stars received, while Panel B shows the t-statistics. Vertical dotted lines in both panels mark the actual coefficient estimate and t-statistics for comparison.

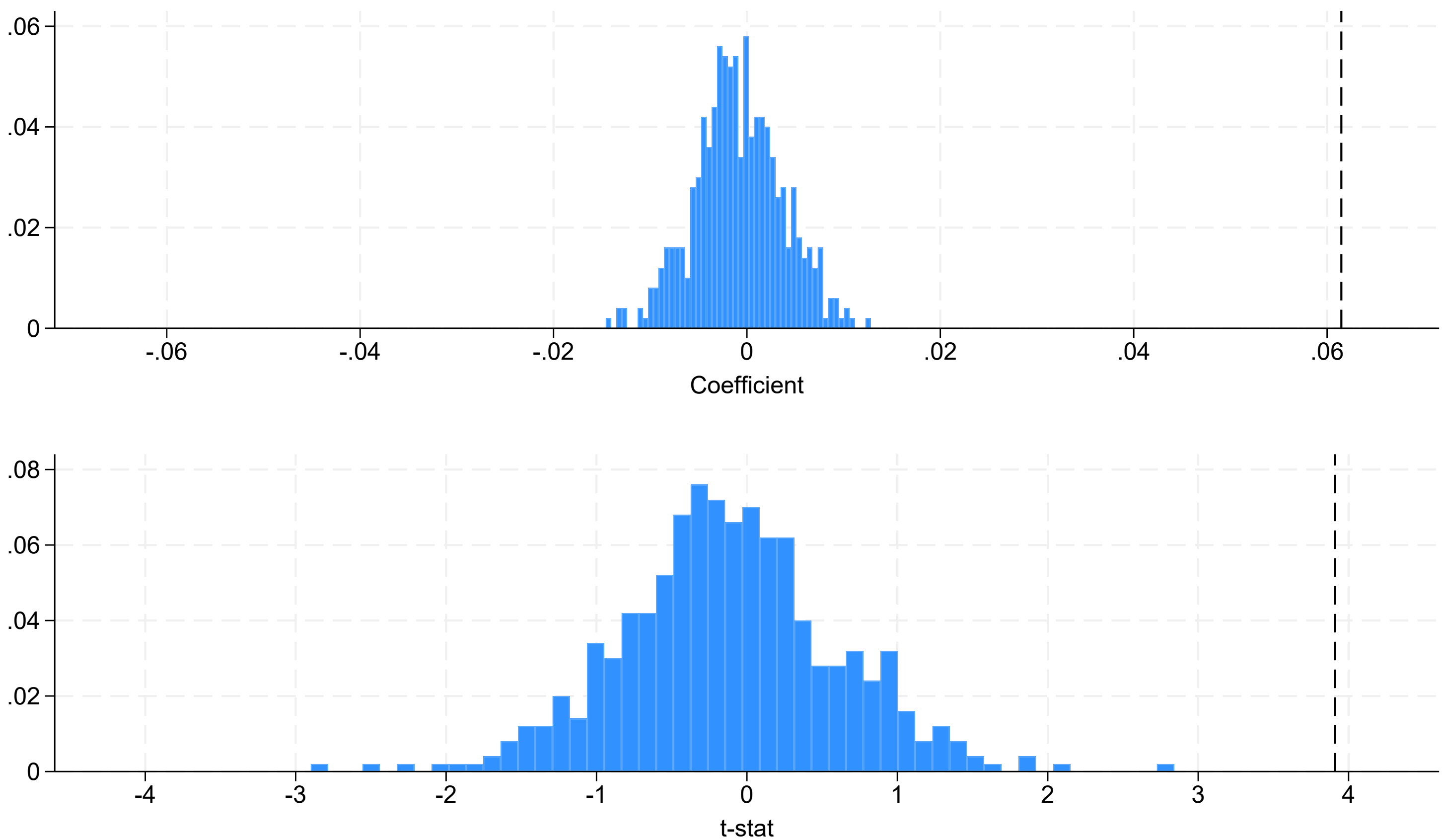

Table 1
Summary of GitHub Activity

Panel A presents the prevalence of GitHub activity among firms, segmented by industry based on the modified Fama-French 5 Industries classification, with an explicit distinction made for the computer software industry. This panel details the percentage of firms engaged in GitHub activities and the distribution of repository ownership within each industry. Panel B compares key financial characteristics for the year 2022 between firms that are active on GitHub and those that are not. See Table A1 for the definition of variables.

Panel A: GitHub Activity

		Number of Repositories										
	% GitHub	Mean	Std	p25	p50	p75	p90	p95	p99	Total	N Firms
Total	18.1%	30.9	430.1	0	0	0	17	62	425	122,971	3,982
Consumer Durables, NonDurables, Wholesale, Retail, and Some Services	19.0%	23.6	393.4	0	0	0	13	36	197	12,757	541
Manufacturing, Energy, and Utilities	7.5%	1.2	6.4	0	0	0	0	6	38	6961	575
Business Equipment, Telephone and Television Transmission (ex. Computer Software)	34.9%	49.1	229.6	0	0	8	87	156	1253	17,726	361
Healthcare, Medical Equipment, and Drugs	6.4%	0.8	5.7	0	0	0	0	2	24	720	861
Other – Mines, Constr, BldMt, Trans, Hotels, Bus Serv, and Entertainment	20.6%	13.3	65.8	0	0	0	14	46	336	5,558	418
Computer Software	62.6%	226.5	1298.7	0	13	82	340	657	4551	82,896	366
Finance	10.8%	3.4	17.3	0	0	0	2	18	93	2,173	636

Panel B: Financial Statistics

	GitHub Firms		Non-GitHub Firms	
	Mean	Median	Mean	Median
Market Capitalization	32,220,913	3,743,408	1,418,477	71,523
Employees	32.5	4.7	7.9	1.1
Number of Patents	1,264	13	67	0
Market-to-Book	6.26	3.49	2.69	1.55
Return-on-Assets	-1.39%	2.26%	-1.35%	3.25%
Investment	3.34%	2.06%	7.06%	4.30%
Annual Returns	14.41%	6.50%	15.21%	5.74%
Sales Growth	14.93%	9.11%	17.27%	8.88%
Tangibility	14.87%	9.02%	27.35%	20.86%
R&D Exp / Total Assets	8.22%	4.79%	3.99%	0.00%
Market Power	3.13	2.21	2.29	1.64
Scope	11	10	8	7
Product Market Centrality	0.0043	0.0024	0.0086	0.0039
Product Market Similarity	4.16	1.74	11.64	2.00
Product Market Fluidity	5.23	4.91	7.67	6.92

Table 2
Determinants of Open-Source Activity

This table reports regression results to examine the factors influencing the extensive and intensive margins of GitHub activity among U.S. public firms. The dependent variable in Columns (1)-(2), *Github*, is a dummy variable that equals one from the first month a firm engages in any open-source activity (open-source firm) on GitHub. In Columns (3)-(4), the dependent variable is the natural logarithm of one plus the number of commits made to repositories owned by the open-source firm each month. See Table A1 for the definition of variables. Standard errors double clustered by industry 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)	0.087***	0.016*	1.127***	0.026
	(0.018)	(0.009)	(0.272)	(0.223)
ln(Employees)	0.003	0.024	-0.092	0.532*
	(0.011)	(0.019)	(0.248)	(0.273)
ln(N Patents + 1)	0.043***	0.017	0.613***	0.035
	(0.012)	(0.042)	(0.138)	(0.249)
Market-to-Book	0.014*	0.002	-0.025	0.081*
	(0.008)	(0.003)	(0.050)	(0.047)
Return-on-Assets	-0.008*	0.002	-0.178***	0.017
	(0.004)	(0.002)	(0.060)	(0.049)
Investment	0.002	-0.000	0.014	0.101
	(0.004)	(0.002)	(0.128)	(0.101)
Return (t-12 to t-1)	-0.007***	-0.003***	-0.072**	-0.041**
	(0.002)	(0.001)	(0.031)	(0.017)
Sales Growth	-0.001	-0.001	0.065	-0.060*
	(0.004)	(0.002)	(0.042)	(0.036)
Tangibility	-0.003	-0.010	0.130	-0.015
	(0.010)	(0.007)	(0.174)	(0.178)
R&D Exp/Total Assets	0.029**	-0.003	0.280*	0.056
	(0.012)	(0.005)	(0.149)	(0.091)
R&D Exp Missing	-0.041**	-0.014	0.416*	-0.777***
	(0.018)	(0.016)	(0.234)	(0.218)
Market Power	0.009	0.001	0.215***	-0.117***
	(0.007)	(0.004)	(0.044)	(0.030)
Scope	-0.005	0.003	0.066	-0.050
	(0.010)	(0.005)	(0.122)	(0.085)
Product Market Centrality	-0.037**	-0.001	-0.472**	0.168*
	(0.016)	(0.015)	(0.213)	(0.092)
Product Market Similarity	0.002	-0.005	0.149	-0.078**
	(0.017)	(0.011)	(0.239)	(0.034)
Product Market Fluidity	0.009	-0.006*	0.120**	-0.079*
	(0.007)	(0.003)	(0.052)	(0.046)
Observations	208,528	208,513	26,422	26,413
Adj. R2	0.331	0.866	0.327	0.781
Industry x Time FE	✓	✓	✓	✓
Firm FE		✓		✓
Sample	All firms	All firms	GitHub = 1	GitHub = 1

Table 3
Summary of Repository Value

This table reports summary statistics of repository values estimated through the methodology detailed in Section 4.1. Panel A summarizes announcement returns and repository values. R is the three-day cumulative market-adjusted announcement return. $E[v|R]$ is the conditional expected return attributable to the repository announcement. ξ is the estimated repository value. "Permissive ξ" is the estimated value of repositories with a permissive open-source license. "Stars," as a measure of popularity, is the number of stars as of February 2024. Panel B provides a breakdown of repository values by industry, adhering to a modified version of the Fama-French 5 Industries classification and distinguishing the computer software industry separately. Panel C lists the top 10 firms based on the aggregate value of their GitHub repository portfolios over the period 2015-2023. Panel D lists the top 10 programming languages based on their aggregate value over the period 2015-2023.

Panel A: Summary Statistics

	Mean	Std	p1	p5	p10	p25	p50	p75	p90	p95	p99	N
R	0.20%	3.52%	-9.13%	-4.54%	-3.03%	-1.28%	0.08%	1.58%	3.78%	5.02%	10.02%	30,034
$E[r\|R]$	0.27%	0.32%	0.00%	0.02%	0.03%	0.06%	0.13%	0.39%	0.68%	0.89%	1.36%	30,034
ξ	832,260	1,044,785	1,646	9,672	23,714	126,456	548,444	1,135,321	1,940,992	2,735,766	5,109,894	30,034
Permissive ξ	880,752	988,449	2,771	15,379	37,234	203,760	643,956	1,185,363	1,954,049	2,684,446	4,711,116	20,897
Stars	208.9	2208.5	0	0	0	2	9	43	200	549	3640	30,034

Panel B: ξ by Industry (in \$)

	Mean	Std	p1	p5	p10	p25	p50	p75	p90	p95	p99	N
Consumer Durables, etc.	1,093,772	871,135	6,630	47,383	317,250	594,828	895,428	1,389,927	1,997,549	2,634,222	4,598,752	7,257
Manufacturing, Energy, and Utilities	274,664	272,437	5,282	8,986	15,390	92,416	214,451	342,402	561,448	834,175	1,420,212	163
Business Eq., etc. (ex. Computer Software)	840,368	1,662,945	1,653	9,363	18,832	23,714	162,723	769,985	2,564,567	4,118,724	8,099,337	3,706
Healthcare, Medical Equipment, and Drugs	533,670	577,298	151	9,240	13,089	25,951	270,544	949,697	1,446,393	1,733,041	1,993,157	97
Other	347,177	415,863	2,681	4,636	7,682	35,599	217,647	459,214	968,190	1,261,965	1,731,836	524
Computer Software	784,464	951,648	1,644	9,297	23,021	129,155	465,133	1,126,993	1,948,435	2,678,231	4,085,496	17,232
Finance	256,762	357,830	718	8,292	11,077	20,100	141,247	283,961	726,312	1,137,260	1,567,406	350

Panel C: Firms with Most Valuable GitHub Portfolios

	Repo Portfolio ξ (in \$)	% of Total ξ	N Repos	N Permissive Repos
Amazon.com Inc	7,839,894,685	31.4%	6,823	6,242
Microsoft Corp	7,785,444,748	31.1%	6,868	4,995
Meta Platforms Inc	2,288,271,347	9.2%	1,170	526
Alphabet Inc	1,769,719,046	7.1%	1,710	1,484
NVIDIA Corporation	1,396,316,951	5.6%	581	288
Apple Inc	1,075,415,551	4.3%	727	135
Salesforce Inc	497,490,305	2.0%	1,039	773
Adobe Inc	225,803,266	0.9%	348	270
International Business Machines Corp	211,951,043	0.8%	613	317
Oracle Corp	206,993,770	0.8%	312	222
...	...	...	...	...
Total	24,996,107,451		30,034	20,897

Panel D: Most Valuable Programming Languages

	Total ξ (in \$)	Mean	Median	N Repos	N Permissive Repos
Python	6,876,648,682	1,157,880	828,729	5,939	4,458
TypeScript	1,930,879,906	837,692	625,554	2,305	1,848
JavaScript	1,744,282,297	554,621	273,981	3,145	2,269
Jupyter Notebook	1,570,756,301	1,306,786	944,270	1,202	960
C#	1,260,655,842	797,379	566,780	1,581	1,135
Java	1,220,972,767	641,942	415,426	1,902	1,421
C++	980,974,664	866,585	524,134	1,132	657
Shell	785,598,748	795,140	557,092	988	723
Go	756,971,189	519,185	228,824	1,458	1,166
HTML	607,608,562	661,883	361,029	918	542

Table 4
Repository Value and Future Popularity

This table reports regression results to validate the value measure for GitHub repositories (the dependent variable, ξ). The key variable of interest is $ln(Stars+1)$. "Stars", as a measure of popularity, is the number of stars as of February 2024. See Table A1 for the definition of variables. Standard errors double clustered by industry 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)	0.207*	0.170*	0.098***	0.087***	0.077***
	(0.102)	(0.086)	(0.024)	(0.025)	(0.019)
ln(Mkt Cap)	1.569***	1.680***	1.748***	1.801***	1.557***
	(0.280)	(0.132)	(0.043)	(0.092)	(0.252)
ln(Volatility)	0.377***	0.537***	0.624***	0.470***	
	(0.075)	(0.108)	(0.045)	(0.043)	
ln(Employees)	-0.091	0.136	0.210***	0.061	
	(0.189)	(0.110)	(0.043)	(0.060)	
ln(Total Patent Value + 1)	0.084	0.047	0.052	0.021	
	(0.064)	(0.058)	(0.041)	(0.107)	
Observations	28,879	28,879	28,879	28,879	28,879
Adj. R2	0.694	0.757	0.811	0.846	0.855
Year FE	✓	✓			
Industry 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), firm characteristics (Panel B), and product market characteristics (Panel C) are correlated with GitHub repository value (ξ). Control variables include market capitalization, volatility, employees and patent value used in Table 4. See Table A1 for the definition of variables. Standard errors double clustered by industry and year 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)
ln(Stars + 1)	0.099***	0.099***	0.110***	0.095***	0.157***	0.156***
	(0.023)	(0.024)	(0.026)	(0.023)	(0.030)	(0.033)
Copyleft License	0.070					0.087
	(0.054)					(0.056)
Other License	0.021					0.012
	(0.025)					(0.025)
Template		-0.244***				-0.226***
		(0.050)				(0.061)
ln(Repo Size + 1)			-0.029*			-0.023
			(0.014)			(0.019)
ln(N Repos + 1)				-0.262***		-0.259***
				(0.040)		(0.041)
ln(N Issues Opened)					-0.080***	-0.072***
					(0.020)	(0.019)
Observations	29,215	29,215	29,215	29,215	29,215	29,215
Adj. R2	0.816	0.816	0.816	0.819	0.816	0.82
Controls	✓	✓	✓	✓	✓	✓
Industry x Year FE	✓	✓	✓	✓	✓	✓

Panel B: Firm Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ln(Stars + 1)	0.098***	0.092**	0.095**	0.098***	0.097**	0.096***	0.094**	0.089**
	(0.025)	(0.028)	(0.029)	(0.025)	(0.030)	(0.028)	(0.033)	(0.033)
Market-to-Book	-0.002							-0.010
	(0.029)							(0.030)
Return-on-Assets		0.083***						0.077**
		(0.019)						(0.024)
Investment			0.045					0.003
			(0.048)					(0.037)
Return (t-1)				0.002				0.004
				(0.013)				(0.017)
Sales Growth					0.046			0.028
					(0.045)			(0.041)
Tangibility						0.040		-0.030
						(0.057)		(0.047)
R&D Exp/Total Assets							0.088	0.089
							(0.075)	(0.061)
R&D Exp Missing							0.473**	0.403***
							(0.142)	(0.099)
Observations	28,074	28,074	28,074	28,074	28,074	28,074	28,074	28,074
Adj. R2	0.804	0.805	0.804	0.804	0.804	0.804	0.805	0.806
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Industry x Year FE	✓	✓	✓	✓	✓	✓	✓	✓

Panel C: Product Market Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)
ln(Stars + 1)	0.080**	0.078**	0.089***	0.082**	0.081**	0.077**
	(0.025)	(0.028)	(0.024)	(0.028)	(0.025)	(0.027)
Market Power	0.102***					0.070**
	(0.019)					(0.022)
Product Market Centrality		0.018				0.124***
		(0.012)				(0.018)
Scope			-0.134***			-0.087**
			(0.034)			(0.032)
Product Market Similarity				-0.078		-0.008
				(0.046)		(0.030)
Product Market Fluidity					-0.144***	-0.122***
					(0.035)	(0.034)
Observations	24,226	24,226	24,226	24,226	24,226	24,226
Adj. R2	0.819	0.817	0.82	0.818	0.819	0.822
Controls	✓	✓	✓	✓	✓	✓
Industry x Year FE	✓	✓	✓	✓	✓	✓

Table 6
Repository Values and Firm Output

This table reports the relation between public repository values (ξ) and a firm's future outcomes over horizons from year $t+1$ to year $t+3$, as formulated in Equation 2. Dependent variables include growth of profits (Panel A), growth of market share (Panel B), growth of patent portfolio value (Panel C), and the natural logarithm of number of employees plus one (Panel D). The left panel looks at the total value of repositories made public in year t by the firm itself, while the right panel looks at the total value of repositories made public in year t by competitors in the same industry (identified by their 3-digit SIC code). The value for each repository is estimated through the methodology detailed in Section 4.1. Control variables include one lag of the dependent variable, market capitalization, volatility, employees, and patent value. See Table A1 for the definition of variables. Standard errors are clustered by firm and year. and 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						
$\ln(\xi+1)$	0.002***	0.006**	0.010***	0.001	0.002	0.002
	(0.001)	(0.002)	(0.004)	(0.001)	(0.003)	(0.002)
Panel B: Profits						
$\ln(\xi+1)$	0.006***	0.009**	0.011***	-0.003	-0.000	0.000
	(0.002)	(0.002)	(0.005)	(0.003)	(0.002)	(0.002)
Panel C: Market Share						
$\ln(\xi+1)$	0.004**	0.009***	0.016***	-0.003	-0.005	-0.006
	(0.002)	(0.003)	(0.003)	(0.003)	(0.005)	(0.007)
Panel D: Labor						
$\ln(\xi+1)$	0.003***	0.005**	0.006	0.000	-0.001	-0.002
	(0.001)	(0.002)	(0.004)	(0.001)	(0.001)	(0.002)
Panel E: Number of Patents						
$\ln(\xi+1)$	0.009**	0.020**	0.031**	0.001	0.003	0.007**
	(0.004)	(0.005)	(0.009)	(0.001)	(0.002)	(0.002)
Panel F: Value of Patents						
$\ln(\xi+1)$	0.005**	0.012**	0.021**	0.001	0.001	0.001
	(0.002)	(0.003)	(0.006)	(0.001)	(0.001)	(0.001)
Controls	✓	✓	✓	✓	✓	✓
Industry x Year FE	✓	✓	✓	✓	✓	✓

Appendices

Appendix A

Table A1
Variable Definitions

Variable	Definition
Copyleft License	Indicator variable that equals one if the repository has a license with some copyleft restrictions (GitHub API).
Employees	Number of employees (Compustat).
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, 2024).
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).
Number of Patents	Number of patents granted (Kogan et al., 2017).
Other License	Indicator variable that equals one if the repository has a customized license that cannot be cleanly classified into "copyleft" or "permissive" categories.
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).
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).
Return (t)	Returns from month t (CRSP).
Return-on-Assets	Net income divided by lagged total assets (Compustat).
Sales	Annual Sales (Compustat).
Scope	Number of industries in which the firm operates, see Hoberg and Phillips (2024) (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).
Volatility	Standard deviation of daily returns over one month (CRSP).
ξ	An estimate of the economic value of repositories (in 2023 dollars) using stock market returns around the repository release date.

Internet Appendix A

Open-Source Licenses

Table IA1
License Classification Based on Permission Levels

This table classifies common licenses of GitHub repositories based on their permission levels. 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. Among open source licenses, permissive licenses impose minimal restrictions, whereas copyleft licenses require that derived works also be open source.

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
Copyleft 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
Other License				
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	
Open source with copyright notice	Can be either permissive or copyleft			

Internet Appendix B

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 announcement of the repository, $[t,\, t+2]$. We choose this window for multiple reasons. First, it is the same window used by prior studies, and so ensures the comparability of our estimates with those for other assets. Second, it limits 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 announcement, the additional noise could render the estimates largely uninformative. Third, we contend that announcements of open-source repositories are predominantly unexpected, so we do not include days prior to the announcement in the announcement window.[IA1]

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 announcement window for repository i, which we label R_i. We assume that R_i is a function of both investor reaction to the repository announcement, v_i, and idiosyncratic noise, ε_i, such that

$$R_i = v_i + \varepsilon_i. \tag{3}$$

We construct the estimate of repository value as the product of the investor reaction to

[IA1] In estimating patent values, Kogan et al. (2017) apply an adjustment for the fact that the announcements are of patent *grants*, while information regarding the patent is first revealed to investors when 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 announcement windows reflect the full value of the repositories.

the repository announcement and the firm's market capitalization on the day prior to the announcement. If multiple repositories are announced on the same day, we assume the value is evenly distributed across those repositories. Given that repository announcements do not follow a typical schedule,[IA2] multiple repository announcements on the same day tend to, anecdotally, correspond to a single project. The value of repository i, ξ_i, is thus calculated as

$$\xi_i = \frac{1}{N_i} E[v_i|R_i] M_i, \tag{4}$$

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 announcement 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 announcement.

To estimate the conditional expected return in Equation (4), we adopt the same distributional assumptions about v and ε as Kogan et al. (2017).[IA3] 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. Under these assumptions, the conditional expected return can be calculated as

$$E[v_i|R_i] = \delta R_i + \sqrt{\delta}\sigma_{\varepsilon ft} \frac{\phi\left(-\sqrt{\delta}\frac{R_j}{\sigma_{\varepsilon ft}}\right)}{1 - \Phi\left(-\sqrt{\delta}\frac{R_j}{\sigma_{\varepsilon ft}}\right)}, \tag{5}$$

where ϕ and Φ represent the standard normal PDF and CDF, respectively, and δ denotes

[IA2] In comparison, patent grants are announced every Tuesday.

[IA3] 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.

the signal-to-noise ratio,

$$\delta = \frac{\sigma^2_{vft}}{\sigma^2_{vft} + \sigma^2_{\varepsilon ft}}. \tag{6}$$

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 announcement window to that of returns over other three-day periods for the same firm within the same year. This comparison takes the regression form

$$\ln(R^2_{fd}) = \gamma I_{fd} + \lambda_{dow} + \eta_{fy} + u_{fd}, \tag{7}$$

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 announcement 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 announcement 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, $\hat{\gamma} = 0.0362$ and $\hat{\delta} = 0.0356$.

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

$$\sigma^2_{\varepsilon ft} = \frac{3\sigma^2_{ft}}{1 + 3d_{ft}(e^{-\hat{\gamma}} - 1)}, \tag{8}$$

where d_{ft} is the fraction of days in the given year that are announcement 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 C

Correlations

This appendix reports univariate correlations among all pairs of variables included in our analysis of the determinants of repository value. The correlation matrix is presented in Table IA2.

Table IA2
Correlation Matrix

This table presents a correlation matrix of all variables included in our analysis of the determinants of repository value. Each variable is defined, along with its data source, in Appendix A1.

	ln(ξ)	ln(Stars + 1)	ln(Mkt Cap)	ln(Volatility)	ln(Employees)	ln(Patent Port ξ + 1)	Permissive License	Other License
ln(Stars + 1)	0.267							
ln(Mkt Cap)	0.774	0.193						
ln(Volatility)	-0.234	-0.172	-0.512					
ln(Employees)	0.659	0.077	0.849	-0.423				
ln(Patent Port ξ + 1)	0.641	0.197	0.840	-0.590	0.767			
Permissive License	0.202	0.100	0.169	-0.053	0.219	0.110		
Other License	-0.189	-0.092	-0.152	0.047	-0.210	-0.099	-0.966	
Template	-0.006	-0.005	0.001	0.011	-0.019	-0.019	-0.001	0.004
ln(Repo Size + 1)	0.023	0.343	0.019	-0.048	-0.023	0.052	-0.028	0.020
ln(N Repos + 1)	0.566	0.049	0.661	-0.313	0.573	0.528	0.203	-0.192
ln(Issues Opened + 1)	0.092	0.729	0.039	-0.150	-0.030	0.080	0.100	-0.096
Market-to-Book	-0.002	-0.128	0.109	0.180	0.018	-0.080	-0.017	0.020
Return-on-Assets	0.457	0.174	0.558	-0.385	0.411	0.619	-0.001	0.007
Investment	0.418	0.062	0.454	0.061	0.630	0.312	0.238	-0.223
Return (t-12 to t-1)	0.099	-0.008	0.209	0.012	-0.003	-0.004	-0.050	0.053
Sales Growth	0.123	0.064	-0.003	0.367	-0.052	-0.261	0.109	-0.096
Tangibility	0.427	-0.003	0.488	0.058	0.713	0.357	0.226	-0.212
R&D Exp/Total Assets	0.112	0.027	0.010	0.423	0.098	-0.112	0.149	-0.137
R&D Exp Missing	-0.180	-0.078	-0.220	0.000	-0.102	-0.221	-0.029	0.016
Market Power	-0.121	0.102	-0.258	0.121	-0.450	-0.198	-0.095	0.098
Scope	-0.011	0.054	-0.070	-0.202	-0.313	-0.018	-0.010	0.012
ln(PM Centrality)	-0.137	-0.004	-0.235	0.013	-0.194	-0.190	0.004	-0.015
PM Similarity	-0.155	-0.006	-0.208	-0.089	-0.365	-0.151	-0.042	0.035
PM Fluidity	-0.057	0.172	-0.118	-0.271	-0.194	-0.024	0.042	-0.044

	Template	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)	-0.019							
ln(N Repos + 1)	0.040	0.011						
ln(Issues Opened + 1)	-0.017	0.374	-0.016					
Market-to-Book	0.003	-0.026	0.055	-0.140				
Return-on-Assets	-0.030	0.053	0.280	0.064	-0.061			
Investment	-0.014	-0.061	0.340	-0.048	0.086	0.155		
Return (t-12 to t-1)	0.011	-0.004	0.071	-0.051	0.333	0.052	-0.006	
Sales Growth	0.007	-0.043	0.031	-0.013	0.277	-0.143	0.394	0.195
Tangibility	-0.008	-0.077	0.355	-0.111	0.123	0.161	0.891	-0.005
R&D Exp/Total Assets	-0.019	-0.026	0.033	-0.035	0.326	-0.227	0.535	0.108
R&D Exp Missing	-0.014	-0.040	-0.327	-0.048	-0.156	-0.063	-0.145	-0.106
Market Power	0.008	0.058	-0.094	0.118	-0.010	0.084	-0.196	-0.117
Scope	0.050	0.080	0.094	0.118	-0.106	0.031	-0.516	0.061
ln(PM Centrality)	-0.004	-0.003	-0.243	0.048	-0.142	-0.217	-0.304	-0.127
PM Similarity	0.022	0.059	-0.027	0.096	-0.097	-0.226	-0.471	-0.056
PM Fluidity	0.000	0.049	-0.125	0.214	-0.229	-0.178	-0.410	-0.073

	Sales Growth	Tangibility	R&D Exp/Total Assets	R&D Exp Missing	Market Power	Scope	ln(PM Centrality)	PM Similarity
Tangibility	0.264							
R&D Exp/Total Assets	0.528	0.491						
R&D Exp Missing	-0.121	-0.073	-0.334					
Market Power	0.142	-0.321	-0.057	-0.107				
Scope	-0.199	-0.547	-0.294	-0.011	0.162			
ln(PM Centrality)	-0.159	-0.229	-0.044	0.262	-0.014	0.290		
PM Similarity	-0.165	-0.544	-0.283	0.018	0.223	0.603	0.385	
PM Fluidity	-0.241	-0.413	-0.151	0.115	0.016	0.569	0.677	0.473