HARVARD | BUSINESS | SCHOOL

Beefing IT up for Your Investor? Open Sourcing and Startup Funding: Evidence from GitHub

Annamaria Conti
Christian Peukert
Maria Roche

Working Paper 22-001

Beefing IT up for Your Investor? Open Sourcing and Startup Funding: Evidence from GitHub

Annamaria Conti
University of Lausanne

Christian Peukert
University of Lausanne

Maria Roche
Harvard Business School

Working Paper 22-001

Funding for this research was provided in part by Harvard Business School and the Swiss National Science Foundation (Project ID: 100013 188998 and 100013 197807).

Beefing IT up for your Investor?
Open Sourcing and Startup Funding: Evidence from GitHub*

Annamaria Conti[1], Christian Peukert[2], and Maria Roche[3]

[1]IE Business School
[2]HEC Lausanne
[3]Harvard Business School

Abstract

We study the participation of nascent firms in open source communities and its implications for attracting funding. To do so, we exploit rich data on 160,065 US startups linking information from Crunchbase to firms' GitHub accounts. Estimating a within-startup model saturated with fixed effects, we show that startups accelerate their activities on the platform as they approach their first financing round. The intensity of their involvement on GitHub declines in the twelve months after. Startups intensify those activities that rely on external technology sources above and beyond the technologies they themselves control. Exploiting a shock that reduced the relative cost of internal collaborations, we provide evidence that startups' decision to integrate external sources of knowledge in their production function hinges on the relative cost vis-à-vis internal collaboration. Applying machine learning to classify GitHub projects, we further unveil that the most prevalent among these external activities are related to software development, data analytics, and integration. Our results indicate that VCs and renowned investors are the most responsive to these activities.

Keywords: Startups, Technology Strategy, GitHub, Machine Learning, Venture Capital

---

*Conti: annamaria.conti@ie.edu, Peukert: christian.peukert@unil.ch, Roche: mroche@hbs.edu. We thank Kimon Protopapas and Ilia Azizi for excellent research assistance and Jorge Guzman for sharing data from Guzman and Li (2021). Many thanks to Bruno Cassiman, Matt Higgins, David Hsu, Harsh Ketkar, Tobias Kretschmer, Rem Koning, Melissa Perri, and Toby Stuart for advice on drafts. This manuscript benefited from many helpful comments provided at the Digital Economy Workshop, the Digital Initiative Workshop, MAD Conference, the Munich Summer Institute, SCECR, the Strategy Science Conference, and the West Coast Research Symposium. We are grateful for the suggestions provided by Karim Lakhani and members of the Laboratory for Innovation Science at the onset of this work, as well as participants of seminars at the CAS Platform Seminar Munich, Copenhagen Business School, Cornell University, HEC Paris, the Intellectual Property & Innovation Virtual seminar series, LUISS Guido Carli University, MPI for Innovation and Competition, NBER Productivity Seminar, the Strategy Unit Seminar at HBS, Universitat Pompeu Fabra, Warwick University, and the Workshop for Entrepreneurial Finance and Innovation. Annamaria Conti and Christian Peukert acknowledge funding from the Swiss National Science Foundation (Project ID: 100013_188998 and 100013_197807). Maria Roche acknowledges funding from the Harvard Business School Division of Research and Faculty Development.

1 Introduction

Open source communities have been increasing in importance as suppliers of knowledge (Dahlander, 2005). Provided how much firms and society rely on open source (Greenstein and Nagle, 2014), understanding the dynamics of open source usage appears critical. Thus far, much of the work examining use of open source has focused on mature firms or those startups that sell open source products or services (Nagle, 2018a). Moreover, many of the studies in this space rely on small scale or qualitative data (Bonaccorsi et al., 2006; Shah, 2006; Stam, 2009) and identify how firms organize for open source (Germonprez et al., 2017) rather than other dynamics around their use.

In this paper, we examine how nascent firms interact with open source communities to raise capital. This is a crucial question given that the financing environment fundamentally shapes strategic choices very early in the life of a new venture (Dushnitsky and Matusik, 2019; Hellmann and Puri, 2002) in the context of nascent firms and its potential implications for attracting funding seems critical. While prior literature stresses the importance of both the founding team (Bernstein et al., 2017; Gompers et al., 2020) and the underlying technology of a venture (Kaplan et al., 2009) to attract funding, technology idea sources still largely remain a black box.

We contribute new empirical findings on the role of open source technology usage among nascent firms in raising funds. Specifically, we use novel data on startups' digital technology activities on the online development and hosting platform GitHub, which has become the open source platform "where the world builds software" (*GitHub.com*). These data have the potential to provide strategy scholars a tool to quantify theory that lies at the heart of the field, such as resources, particularly knowledge, and the real-time assessment of the recombination thereof. In this paper, we provide an example of how these data can be put to work by examining how startups utilize this platform at around the time they raise a funding round, the type of activities in which they engage, and their sources of knowledge. To do so, we exploit data encompassing 160,065 US startups listed on Crunchbase that were founded

between 2005 and 2020 as our initial sample. We then match firms to organization accounts on GitHub and access their public GitHub records logged since 2011 from the GHArchive. The combination of these two datasets provides us with information on the industry, investors, and total amount of funds a firm raises as well as the type and nature of activities a firm engages in on GitHub.

As a first step, we examine the type of startups that opt for using GitHub. Here, we find a strong positive correlation between different quality measures – funding and human capital – and the likelihood of having a GitHub organization account. Moreover, and perhaps not surprisingly, startups that operate in software-related industries, and the more successful among these, are most likely to have a GitHub account.

Provided there is a notable selection into the open source platform, we choose to focus on this highly relevant sub-population: startups that are on GitHub and raise a financing round at some point. We then use our rich panel data to estimate a within-startup model that relates the dynamics of open source technology strategy, as measured by GitHub activity, to achieving funding milestones. To address omitted variable concerns, we saturate the model with a wide range of fixed effects pertaining to the firm, firm-age, industry-year, region-year, and lead-investor-year. We show that GitHub activity takes off in the twelve months preceding a first funding round, accelerating as the startup approaches the round, and then levels off in the twelve months after. Specifically, our data reveal that an extra twelve months towards raising a first round increases the probability of being active on GitHub by 50% relative to the mean. This effect does not appear to be driven by a startup's general technology life cycle. These results are also robust to different specifications such as a difference-in-differences model comparing the dynamics of GitHub activities at around the time funded startups raise a first financing round to the GitHub dynamics of unfunded startups during the same time period.

Additionally, we find that these dynamics are particularly strong when the first round of funding is raised and when considering seed rounds. Moreover, these dynamics appear to

be most pronounced when funds are raised by VCs and successful investors. Differences in technology and in the strategic positioning relative to competitors seem to play a role.

Strikingly, our results indicate a significant leveling off the likelihood of choosing a permissive license – a license with only minimal restrictions on how software can be used, modified, and redistributed – after a startup raises a first financing round. This result suggests that startups factor in the downsides of relying on open source communities and may react accordingly.

We next unveil crucial heterogeneity among different activities on GitHub. Specifically, we find that startups appear to intensify activities related to external technology sources, that is, repositories they do not control themselves, more so than activities related to internal sources. We extend this analysis by assessing how a startup's reliance on external technology sources depends on the relative cost of accessing external code versus collaborating internally using GitHub. For this scope, we examine a change in GitHub's pricing model that is exogenous to a startup's technology, quality, or competitive position. This change was introduced in October 2015 and led to a substantial decline in the costs of internal collaborations on the GitHub platform, while leaving the cost of accessing external repositories unchanged. With the new pricing model, we observe that startups rely relatively less on external technology sources and, instead, intensify internal collaborations on GitHub compared to the pre-pricing-change period.

To delve deeper into our findings, we make use of natural language processing methods (Miric et al., 2022). We classify external activities on GitHub by the following use-cases: Software Development/Backend (SD/BE), Machine Learning (ML), Application Programming Interfaces (API), and User Interfaces (UI). We find that all use-cases but UI intensify prior to receiving a funding round and decelerate afterwards. While SD/BE, ML, and API might be considered as core technology components, UI serves as an outward facing building block.

We further examine whether startups *merely* engage with GitHub to increase the visibility of their technologies before raising a round. In contrast with this interpretation, we find

that the dynamics of startups making their existing private repositories public do not change before and after the first round. Additionally, the likelihood of publishing repositories that are eventually re-used by other account owners increases as a startup approaches its first financing round and its slope remains unchanged thereafter, suggesting that the trajectory of publishing relevant repositories is similar in the period pre- and post-funding event. On the whole, our results support the notion that startups are "beefing up" their technologies before they receive early-stage financing – by accessing external knowledge from open source communities.

The remainder of the paper is organized as follows. Section II describes the context and the data. Sections III and IV present the empirical specification and the results, respectively. Section V provides an overview of potential applications of these data and Section VI discusses potential interpretations of our findings as well as contributions to the literature and concludes.

2 Empirical setting and data

2.1 GitHub - "Where the world builds software"

GitHub is a hosting service for software development and collaborative version control. With 40 million public repositories in April 2021, GitHub is the largest host of source code[1] and has come to be known as the place "where the world builds software" (*GitHub.com*). Individuals and organizations use GitHub to improve and upgrade their own projects by accessing code and information from other existing projects as well as to contribute to other projects. GitHub has a history of being backed by a number of high-profile investors (e.g. through a $100m investment by Andreesen & Horowitz) and was acquired by Microsoft for $7.5b in 2018.[2]

GitHub offers personal user as well as organization accounts, the latter being the object of our analysis. The source code on GitHub is organized in *repositories*, that is, folders

[1] See `https://GitHub.com/search?q=is:public`, accessed April 25, 2021.

[2] See `https://techcrunch.com/2018/06/04/microsoft-has-acquired-GitHub-for-7-5b-in-microsoft-stock/`, accessed April 25, 2021.

containing projects. The version control system Git allows users to save snapshots of files within a repository. When a user issues a *commit*, a snapshot of the file's contents is created and associated with a timestamp. The repository owners can add members, that is, GitHub user accounts, to their repositories and grant permissions to add content. In our analysis, we will use the addition of members to internally controlled repositories as a measure of increased internal collaboration. The most frequently used way to interact with the repositories of other users is called *forking*. With forking, a user makes a copy of another user's repository, which is then integrated into the initial user's account. Forked repositories can be the foundation for further internal development.

GitHub's pricing model is based on subscription fees. Initially, the subscription price included ten repositories per organization account. Adding additional internal repositories was possible but nearly doubled the monthly subscription fee. Effective October 2015, GitHub introduced a new pricing plan shifting from a repository to a user-based model. While the latter model allowed creating unlimited repositories, the fixed subscription did not increase markedly. For organization accounts, the new price policy implied that it became drastically cheaper to start internal repositories and add members to multiple repositories, thereby lowering the coordination costs associated with using GitHub as a development platform for internal collaboration.[3] We will make use of this sharp discontinuity in the analysis below.

2.2 Data

To build our dataset, we combine data on US startups and their investors, which are available on Crunchbase, with information on their respective GitHub activities available from GHArchive and through the GitHub API.

2.2.1 Crunchbase

Crunchbase serves as our first source of information on startups. This online directory records fine-grained information on the startups, their founders, and their investors. As

[3]For more information refer to https://docs.github.com/en/billing/managing-billing-for-your-github-account/about-per-user-pricing and https://www.infoworld.com/article/3069275/github-ushers-in-unlimited-private-repositories.html. Accessed March 4, 2022.

described by Conti and Roche (2021), a considerable portion of the data are entered by Crunchbase staff, while the remaining part is crowdsourced. Registered members can enter information to the database, which the Crunchbase staff successively reviews. Relative to databases such as VentureXpert and VentureSource, Crunchbase has the advantage of providing larger coverage of technology startups as it also encompasses startups that did not raise venture capital. From Crunchbase, we extracted information pertaining to all the recorded US startups that were founded between 2005 and 2020. This amounts to 160,065 startups, for which we have data encompassing their founding dates, industry keywords, location, financing rounds and participating investors, as well as exit outcomes.

As shown in Table 1a, approximately half of the startups (46%) are located in California, Massachusetts, and New York, reflecting the comparative advantage of these regions in entrepreneurship. Thirty-six percent of them raised at least one round of financing. Additionally, 8% of the startups were acquired as of December 2020 and 1.2% went public through an IPO.

While Crunchbase does not categorize startups into sectors, it provides industry group information for each of them.[4] There are approximately 40 distinct industry groups, and, on average, a startup is assigned three industry group keywords. Using this information, we computed a variable to measure the relatedness of a startup's technology to software. This index is defined as the share of a startup's industry groups that are related to software. The groups related to software are: Apps, Artificial Intelligence, Consumer Electronics, Data and Analytics, Design, Financial Services, Gaming, Information Technology, Internet Services, Messaging and Telecommunications, Mobile, Payments, Platforms, Privacy and Security, and Software. As shown in Table 1a, the mean of this index is 0.46.

⟨ Insert Table 1 about here ⟩

[4]The full list of Crunchbase industry groups is available at https://support.crunchbase.com/hc/en-us/articles/360043146954-What-Industries-are-included-in-Crunchbase-. Accessed March 4, 2022.

2.2.2 GitHub Activities

We use the GitHub API to collect all organization accounts on GitHub. From these accounts, we extract the websites of the account owners. We used this information to link GitHub organization accounts to 14,881 Crunchbase company profiles. Note that over 60% of the startups with a GitHub organization account are described by the software industry group keyword, and also our software share index is higher for startups with GitHub accounts. We further gather time-variant information on the public events of all startups through GHArchive, which is a non-profit service that provides a full record of the public timeline on GitHub since February 2011. This archive includes, among others, time-stamped data on events such as commits and forks – activities related to all own or external public repositories with which an organization account interacts. Furthermore, we observe events related to the addition of new members to repositories, and the creation of new public repositories or making a private repository publicly visible.

Descriptive statistics reported in Table 1b show the top three events through which startups engage with external and internal repositories. In the case of external repositories, the share of forks is very large (96%). In contrast, contributions to external repositories in the form of pushing commits are rare (1%). This distribution provides some evidence that startups access external repositories to upgrade their technologies and not just to provide comments or contribute to other users' projects. As for the distribution of internal events, it appears to be relatively more spread out. The most popular events concern adding members to internal repositories (46%) and making repositories public (17%), but pushing commits are also frequent (15%).

We further employed the GitHub API to collect meta information on all public repositories in a startup's organization account.[5] Building on these data, we distinguish between a startup's public activities related to its own repositories and its engagement with external repositories. As reported in Figure 1, we observe a stark increase over time in both the

[5]Note that account-specific meta-information is time-invariant. We collected it through separate crawls in October 2020, January 2021, and March 2021.

number of public activities related to the startups' own repositories and the number of engagements with external repositories. The vertical line indicates the time of the pricing change, which we exploit in a later section of the paper.

⟨ Insert Figure 1 about here ⟩

As suggested by most recent work, machine learning (ML) methods are powerful in detecting patterns in large and complex data (Choudhury et al., 2021; Miric et al., 2022). In this paper, we use supervised and unsupervised machine learning methods which we describe in detail in the Appendix, to classify the public repositories of all organizations, as well as the external repositories with which the organizations interacted through commits, pull requests or forks according to their type. We distinguish between repositories that pertain to software development/back end (SD/BE), machine learning (ML), application programming interface (API), and user interface (UI). By doing so, we consider a comprehensive set of use-cases that are relevant for the development of a digital technology (Yoo et al., 2012).[6] Because founders increasingly rely on GitHub to upgrade their human resources, we generate an additional category encompassing public repositories related to best practices on HR management (Jain, 2018). For example, these include guidelines for coding interviews and templates for company policies on issues such as working from home, diversity, equity, and inclusion.

We report a descriptive representation of the output obtained from the algorithm in Figure 2. Here, we display the most common words for each of the categories we consider.[7] Figure 3, instead, captures the relevance of each use-case across industry groups. As shown, there is variation across industry groups in the relevance of the different use-cases we consider.

[6]Forking, documenting code, reporting issues through comments and receiving notifications about project changes (as a "watcher") are integral parts of software development on GitHub. A survey among software developers in Jiang et al. (2017) shows that developers predominately fork repositories to submit pull requests, that is, contributions to the codebase of a given repository, fix bugs, add features, and keep copies in case the original owner deletes the repository. As such, forking can be considered a software development tool and not just a means to access off-the-shelf software (Sheoran et al., 2014).

[7]We collect additional data on startups' web technology usage using the *Wappalyzer* technology profiler accessed through the HTTPArchive. We then manually classified the 50 most used web technologies with respect to whether they are open source and which of the use-cases – SD/BE, ML, API, UI – they fall under. This exercise shows that the web technologies startups rely on are 70% open source, and cover the entire range of use-cases we observe on GitHub.

For instance, the share of commits, pull requests, and forks related to UI repositories is smallest for startups in privacy/security and highest in more consumer-facing internet services. Startups in software/IT have the highest share of interactions with API-related repositories, but the lowest share with productivity-related repositories. This suggests that firm- and industry-specific unobservables are important factors to control for in our econometric models.

⟨ Insert Figures 2 and 3 about here ⟩

3 Empirical specification

We begin our empirical investigation by descriptively assessing the correlation between a startup having an organization account on GitHub – which is our proxy for a startup's involvement in an open source community – and attracting funds, from VCs and other investors. We also evaluate how this correlation compares to the correlation between having a public account on GitHub and a startup's human capital – as defined by whether a startup's founder or CXO[8] is highly ranked on Crunchbase's list of top people.[9] To achieve these goals, we estimate the following linear probability model:

$$GitHub_{ifjs} = \alpha + \beta_i Funded_i + \gamma_i TopTeam_i + \eta_f + \nu_j + \psi_s + \varepsilon_{ifjs}, \tag{1}$$

where $GitHub_{ifjs}$ is an indicator that takes the value one if a startup i, founded in year f, developing a technology in industry group j, and located in region s, had an organization account on GitHub as of January 2021 and zero otherwise. The variable $Funded_i$ is an indicator that takes the value one if a startup raised at least one financing round as of January 2021 and zero otherwise, while $TopTeam_i$ is an indicator identifying prominent founders and CXOs. The latter measure equals one if an employee is ranked among the top 1000 by Crunchbase and zero otherwise. In this regression, we include fixed effects for a startup's founding year (η_f) and for whether the startup is located in either Massachusetts, New York,

[8]By CXOs we refer to Chief Executive Officers (CEOs), Chief Technology Officers (CTOs), Chief Financial Officers (CFOs), and Chief Marketing Officers (CMOs).

[9]Refer to: `https://www.crunchbase.com/discover/people`, accessed March 4, 2022.

California, or other states (ψ_s). We additionally include industry group fixed effects (ν_j). The industry groups we consider are Information Technology, Software, Data Analytics, Internet Services, and Artificial Intelligence. As mentioned earlier, a startup can be described by more than one industry group. The results remain similar when we add additional keywords.

In the second part of the empirical analysis, we evaluate the dynamics of a startup's involvement on GitHub at around the time the startup raises a financing round. For this purpose, we restrict the sample to the 10,514 startups that raised at least one round of financing and for which we could identify a public GitHub organization account. Table 1c reports descriptive statistics for this sample. We then estimate a within-startup model, where we assess how the probability that a startup engages in public activity on GitHub in month t varies during the twelve months preceding and succeeding a startup's financing round. Specifically, the main equation we estimate is:

$$Y_{itjs} = \alpha + \beta_{i,t} Post_{i,t} + \gamma_{it}\tau_{it} + \delta_{it} Post_{i,t} \times \tau_{it} + \omega_i + \mu_{it} + \nu_{jt} + \psi_{st} + \rho_{it} + \varepsilon_{itjs}, \quad (2)$$

where Y_{itjs} is an indicator that takes the value one if a startup i, developing a technology in industry group j and located in region s, engages in at least one public activity – across own and external repositories – on GitHub in month t and zero otherwise. In extensions to our baseline analyses, we examine variants of this outcome to better understand the type of GitHub activities in which a startup may engage. The indicator $Post_{it}$ is equal to one for the twelve months that follow a startup's given financing round, and zero in the twelve months preceding it. While the focus of our analysis is on a startup's first financing round, we will also examine the activities of startups as they relate to subsequent financing rounds. The variable τ_{it} is the count of months to/from a given round. The coefficients of interest are γ, which represents the effect of an extra month towards a given financing round on the likelihood that a startup engages in a public activity on GitHub, and δ which represents the effect of an extra month from a given financing round. We cluster standard errors at the

startup level.

By estimating Eq. (2), our goal is to assess the existence of a simultaneous Perfect Bayesian Equilibrium whereby a startup's decision to become involved with the GitHub platform incorporates the expectation of attracting funding, and the investors' decision to fund a startup depends on the startup's involvement with the GitHub platform. In doing so, we would ideally be able to isolate the dynamics of engagement on GitHub from other confounding factors such as a startup's technology and its positioning vis-à-vis competitors, the technology life cycle, and technology shocks. This is crucial given that these simultaneously occurring features may be responsible for any relationship we detect. To get as close as possible to an ideal experiment, we saturate the model with a fine-grained set of relevant fixed effects. In particular, ω_i is a focal startup's fixed effect, which controls for invariant differences across our sample companies. For instance, it is possible that some startups are intrinsically more likely to rely on GitHub, given the characteristics of the technologies they develop. A startup's age (measured in years) fixed effect is denoted by μ_{it}. The inclusion of this fixed effect addresses the concern that any variation in the engagement of a startup on GitHub may be driven by a startup's position in the life cycle rather than by its intent to attract funding. In a more stringent specification, we add age times startup and $Post_{it}$ times startup fixed effects given that life cycle effects may vary by company. We additionally include industry group times year fixed effects (ν_{jt}) to mitigate the concern that the coefficients of interest may be confounded by technology shocks, which could happen at around the time a startup raises a financing round. Moreover, we include location times year fixed effects (ψ_{st}), distinguishing between California, Massachusetts, and New York (that is, the US technology and entrepreneurial hubs) and all other states. The rationale is that possible technology shocks happening in any given period should originate from these hubs. Finally, we include fixed effects for whether a startup raised a subsequent round in any of the t months following the observed round (ρ_{it}). In a more stringent specification, we interact ρ_{it} with τ_{it} to address the possibility that a startup's involvement on GitHub might be driven by the expectation of

attracting subsequent rounds.

4 Results

This section explores (i) the startups' characteristics correlated with having a GitHub account as well as (ii) the dynamics of startups' participation in GitHub before and after raising a financing round.

4.1 Having an organization account on GitHub

The results from estimating Eq. (1) are reported in Table 2. For this analysis, we consider the full set of 160,065 startups. We cluster standard errors by founding year. The dependent variable is the likelihood of having an organization account on GitHub as of January 2021. As shown in column (1), startups that received funding are 8 percentage points more likely to have a GitHub account (p-value: 0.000). As displayed in column (2), the magnitude of this correlation is smaller than the one between having a prominent founder or CXO and having a GitHub account (33 percentage points, p-value: 0.000).[10] In column (3), we replace industry group fixed effects with the share of industry groups that are related to software. Here, we show that the more startups develop software technologies, the more likely they are to have a GitHub account (p-value: 0.000). This result is to be expected given that GitHub is a platform for software development. Finally, in column (4), we show that the correlation between having raised funds and having a GitHub account intensifies with the share of a startup's keywords related to software (p-value: 0.000).

Overall, these correlations highlight that startup quality, measured by either having obtained funds or the level of a startup's human capital, is a consequential factor explaining a startup's involvement on GitHub. Moreover, the positive correlation between a startup's quality and the probability that the firm has a public GitHub account is strongest for startups specialized in software technologies. Our findings here are in line with existing work which suggests that a firms' openness regarding knowledge flows may relate to a firms' initial capabilities and endowments (Greul et al., 2018).

[10] In Table A1, we adopt less stringent definitions of $TopTeam_i$ showing that the main results remain invariant.

⟨ Insert Table 2 about here ⟩

4.2 Startup GitHub activities before and after raising a financing round

In this section, we examine how a startup's activities on GitHub vary before and after a given financing round. We begin by reporting the main results and successively explore the potential mechanisms driving them.

4.2.1 Main results

How do startups engage with the open source community on Github when raising a financing round? It is possible that startups may accelerate their contributions on GitHub to develop their technologies in order to attract investors. Yet, such engagement is not exempt from drawbacks, which include appropriability threats and coordination costs (David and Greenstein, 1990; Greenstein, 1996). In what follows we will examine these dynamics, and this potential trade-off in close detail.

To assess startup open source strategies, we estimate Eq. (2) for the sample of 10,514 startups that raised at least one financing round and have an organization account on GitHub. We focus on this sample given our earlier finding that GitHub account owners tend to be the most successful startups, and thus, it is difficult to find an appropriate control group for these ventures. During the twelve months until the first financing round, our model compares startups that have yet to raise a financing round with those that have already raised one, holding fixed startup characteristics constant. Vice-versa, during the twelve months succeeding the first financing round, our model compares startups that have raised a financing round with those that have yet to raise one.[11] We report the results on Table 3 and in Figure 4. Here, we specifically focus on a funded startup's first financing round. In the next section, we extend the analysis to a startup's subsequent rounds.

Column (1) of Table 3 includes startup and year fixed effects in addition to the fixed

[11] Given our sample definition, Eq. (2) does not estimate an average treatment effect (ATE) for the population of firms observed on Crunchbase, but rather an average treatment effect on the treated (ATT). This is perhaps the most relevant subsample given the contributions these startups make to employment and innovation. It is, however, likely that our ATT is larger than the ATE.

effects for whether a startup raised successive rounds in any of the twelve months following the first round (ρ_{it}). We further control for a startup's position in the life cycle by including the natural logarithm of a startup's age. The positive coefficient associated with the indicator $Post_{it}$ suggests that, all else equal, startups become more active on GitHub after they raise their first financing round (coefficient magnitude: 0.042, p-value: 0.000). The coefficient associated with the time trend τ_{it} is positive, indicating that the likelihood that a startup engages in a public activity on GitHub increases as the startup gets closer to raising its first round. The magnitude of the coefficient can be interpreted such that an extra month in a startup's life cycle is associated with a 0.34 percentage points increment in the likelihood that a startup engages in a public activity on GitHub (p-value: 0.000). Compounding over a year, this effect represents a roughly 50% increase relative to the outcome mean. Moreover, the negative coefficient of the interaction between $Post_{it}$ and τ_{it} suggests that a startup's participation in GitHub becomes less intensive the further in time the startup is from the first round. The magnitude of the coefficient suggests that an extra month away from a startup's first round reduces the positive slope of the probability that a startup engages in a public activity on GitHub by 0.2 percentage points (p-value: 0.000).

In column (2) of Table 3, we add region by year and industry group by year fixed effects. As shown, the magnitudes of the coefficients and the related p-values remain very similar to those displayed in column (1). In column (3), we replace our industry group keywords with the share of a startup's industry group keywords that are related to software. The results remain essentially unchanged when we interact the newly generated variable with year fixed effects.

In column (4), we estimate a similar model as the one in column (2), this time replacing the natural logarithm of a startup's age with age fixed effects to more flexibly control for life cycle effects. We also add interactions between ρ_{it} and τ_{it}. This specification, which is our preferred one, delivers very similar results to those reported in columns (1) to (3). In column (5), we estimate a more stringent specification than in column (4), allowing for the

possibility that life cycle effects may vary by startup and by whether the startup has already raised a first round. For this purpose, we include age times startup and $Post_{it}$ times startup fixed effects. The magnitudes of the effects are stronger than those displayed in the previous columns (the related p-values remain unchanged) and the results continue to show that a startup's engagement on GitHub intensifies as a startup approaches their first round and levels off afterwards.

Finally, in column (6) of Table 3, we add lead investor by year fixed effects to the specification in column (5), thus controlling for shocks that may be common to startups financed by the same investor. To identify lead investors, we employed the categorization of lead investors provided by Crunchbase. When this was missing, we considered the investor who backed the highest number of a focal startup's financing rounds as the lead investor. By adopting this procedure, we identified lead investors for 62% of the startups. Reassuringly, the results continue to hold.[12]

⟨ Insert Table 3 about here ⟩

In Figure 4 we visually depict the results provided in the regression table. In particular, we report the results of an event study where we replace τ_{it} with dummies for each of the months preceding and following a first financing round. We include the same controls and fixed effects as in column 4 of Table 3. The figure displays a non-linear relationship, with a clear uptick in activities just before receiving financing proceeded by a leveling off right after.

⟨ Insert Figure 4 about here ⟩

The estimates presented so far are for the subset of startups with a GitHub account and that raised at least one financing round. Yet, one may want to assess whether and how these effects would vary if we were to compare funded startups with a similar set of control startups that did not attract capital. We explore this venue in Table A3 where we estimate

[12] In Table A2, we modify the sample size and only retain startups that opened a GitHub account at least twelve months prior to raising their first financing round. The results remain qualitatively invariant.

a difference-in-differences model comparing the dynamics of GitHub activities at around the time funded startups raise a first financing round to the GitHub dynamics of unfunded startups during the same time period. Control startups are randomly chosen from the set of startups that were founded during the same year and in the same state as the treated startups, and had a similar top team structure and share of software keywords. Reassuringly, the results in column 1, where we include the same set of fixed effects as in column 5 of Table 4 in addition to fixed effects for groups of treated-control startups, show that GitHub activities of treated startups accelerate more than activities by untreated startups before treated startups raise their first round. An extra month towards raising a first round is associated with a 0.04 percentage point increase in the likelihood that a startup engages in a public activity on GitHub. The effect size is similar as the one displayed in column 5 of Table 4 for τ_{it}. On the contrary, the effect for untreated startups is only 0.01 percentage points. Moreover, a month away from a startup's first round is associated with a 0.03 percentage points decrease in the likelihood that the startup engages in a public activity on GitHub. This effect is again similar to the one reported in column (5) of Table 4 for $Post_{it}$ times τ_{it}. Vice-versa, the likelihood that untreated startups engage in a public activity on GitHub does not change after treated startups raise their first round. Overall, these results provide an indication that the estimates we display in Table 4 apply not only the subset of startups that eventually raised a financing round but also to comparable unfunded startups.

While the findings presented so far suggest that – all else equal – startups accelerate their participation on GitHub to develop their technologies and attract funds, it is possible that such participation reflects a startup's technology and technology life cycle rather than its reliance on an open source community. The inclusion of startup and startup by age fixed effects should at least in part control for a startup's technology aspects, but some of them may remain unaccounted for. To further address this concern, we delve deeper into the type of GitHub activities in which a startup engages. For this purpose, we modify the dependent variable in Eq. (2) and examine whether a startup engages in public activities

related to its own repositories. This measure proxies for a startup's internal technology investments. Additionally, we investigate whether a startup engages with external repositories (mostly through forking). Since a startup does not have direct control over these repositories, this additional outcome provides an indication of whether a startup builds on repositories controlled by other GitHub users to develop its technologies. If our main results were to capture the life cycle of a startup's technology only, we would observe similar trends for both types of activities.

The distinction we make between engagements with internal and external repositories builds on the analysis we provide in Table A4, where we zoom in on the different types of GitHub activities and show that the dynamics of forking external repositories are quite different from those of other activities at around the time of a startup's first financing round. As forking a repository from an external account means that the externally sourced code becomes available for internal use, this outcome epitomizes the strategy of integrating external knowledge from the open source community. Such openness towards integrating external knowledge has been suggested to generate important learning effects enabling firms to improve their innovative output (Love et al., 2014).

The results from this analysis are reported in Table 4 and in Figure 5. They show that, prior to raising a first round, startups accelerate their engagement with external repositories (coefficient magnitude of τ_{it}: 0.003; p-value: 0.000) more than they do for their internal repositories (coefficient magnitude of τ_{it}: 0.001; p-value: 0.000). These results corroborate our main findings and suggest that reliance on external knowledge made available by the open source community is potentially more important than internally developed knowledge to attract investments.[13]

Supplementary evidence provided in Table A6 shows that the dynamics of a startup's engagement with external repositories prior to raising a first financing round are not driven by

[13]In Table A5, we find that the results remain similar when we use the *share* of external activities as an outcome. With this specification we de-trend the engagement of a startup with external repositories, accounting for confounding factors such as a startup's technology trajectory and variations in the positioning of a startup vis-à-vis competitors and in hiring policies.

the forking of repositories with permissive licenses, that is, licenses that allow for commercial re-use. Therefore, our findings may be interpreted such that when startups prepare for raising their first financing round, they do not simply engage in the repackaging and the "resale" of technology produced by others.

⟨ Insert Table 4 and Figure 5 about here ⟩

We further shed light on the drivers of a startup's reliance on external technology sources by assessing how GitHub activities vary depending on their relative cost. For this purpose, we exploit a change in GitHub's pricing model which occurred in October 2015. While under the old pricing model account owners had to pay per repository, the new model introduced user-based pricing. Upon paying a subscription, organizations would now be able to create an unlimited number of repositories allowing their members to more efficiently organize their internal collaborations on GitHub. While the marginal costs of internal collaborations declined, the new pricing scheme left the marginal cost of interacting with external repositories unchanged. As a result, the *relative* cost of internal collaborations declined. We measure internal collaborations using "Member Events", i.e. the addition of new collaborators to existing internal repositories. As displayed in Figure 6, the decline in the relative cost of internal collaborations had a considerable positive impact on the extent of such internal collaborations, starting from the moment the new pricing model was implemented. Conversely, startups' activities related to external repositories remained unchanged.

⟨ Insert Figure 6 about here ⟩

The important feature of this pricing reform is that it is exogenous to a startup's technology life cycle. Therefore, we can assess how exogenous variation in the relative cost of accessing external sources of knowledge affects the startups' willingness to rely on an open source community to develop their technologies and attract funds. For this scope, we modify Eq. (2) to include the indicator $NewPriceScheme_t$ which identifies the period following the

introduction of the new pricing scheme. We additionally introduce interaction terms between $NewPriceScheme_t$ and $Post_{it}$, τ_{it}, and $Post_{it} \times \tau_{it}$.

The results are reported in Table 5 and Figure 7. The dependent variables we consider are indicators for whether a startup engages in a member event in month t (column (1)), for whether a startup engages in a non-member internal event (column (2)), and for whether a startup engages with an external repository (column 3). We include the full set of fixed effects as specified in Eq. (2).[14]

As shown, prior to the pricing reform, the engagement of startups with external repositories increases when startups are on the verge of raising their first financing round and subsequently levels off. Conversely, the likelihood of adding members to internal repositories is close to zero and invariant in the months preceding and following the financing round. After the pricing reform, we observe that firms accelerate their internal collaborations, but less so their engagement with external repositories, in the months preceding their first financing round compared to the pre-pricing-change period. Overall, these findings suggest that a startup's decision to integrate external sources of knowledge from an open source platform in the production of its technologies crucially depends on their relative cost: startups opt for these external sources when the cost of accessing them is comparatively low. In line with previous theoretical work, it seems that firms may pursue more open source in order to improve their performance by reducing their base costs (Alexy et al., 2018).

⟨ Insert Table 5 and Figure 7 about here ⟩

Having demonstrated that startups accelerate their engagement with open source communities prior to attracting funds, we delve deeper into the trade-off between open source and appropriability by assessing changes in startups' licensing strategies. Licensing decisions have been suggested to be critical governance decisions and of utmost importance to open source communities (He et al., 2020). Moreover, prior work attributes a large role to the availability

[14]Results with the full difference-in-differences specification comparing how the new pricing scheme affected external versus internal events are reported in Table A7. They are very similar to those displayed in Table 5.

of public data in predicting firm performance outcomes (Nagaraj, 2022).

Building on and extending this work, we examine the types of licenses that startups choose for their new repositories before and after raising a financing round. In particular, we distinguish between more or less permissive strategies based on whether or not they grant use rights, including the right to re-license.[15] The results from this analysis are reported in Table 6. Here, we show a leveling off of a startup's likelihood of adopting permissive licenses after the first round (largest p-value associated with $Post_{it} \times \tau_{it}$: 0.007). This result suggests that startups may factor in the downsides of relying on open source communities and less restrictive intellectual property protection reacting accordingly once funding has been secured.

⟨ Insert Table 6 ⟩

We next examine the heterogeneity of a startup's activity on GitHub, categorizing use-cases of repositories according to whether they pertain to SD/BE, ML, API, and UI. By doing so, we consider fundamental aspects related to the production of a digital technology, encompassing the back end, data analysis, the front end, and the interconnection between front end and back end. Further, founders increasingly rely on GitHub for activities related to human resources, we additionally examine whether a startup resorts to GitHub to access best practices on HR management.

The results are reported in Table 7. In Panel A of this table, we focus on public GitHub activities related to a startup's own repositories, while in Panel B we examine a startup's engagement with external repositories. In column (1) we show that, prior to raising a first round, startups intensify both their internal activities and their engagement with external repositories related to SD/BE. However, while a startup's external engagement significantly fades after raising a first round, the increment in the likelihood of investing in SD/BE internally does not significantly change after raising a first round. In column (2), we show

[15] Examples of permissive licenses are BSD, MIT, Apache, and CC-BY. For a comprehensive list, refer to https://gist.github.com/nicolasdao/a7adda51f2f185e8d2700e1573d8a633. Accessed March 4, 2022.

that a startup's investment in ML prior to raising a first round rests on a startup's engagement with external repositories. In column (3), where we consider a startup's investment in API, we find a steady increase in the likelihood that a startup engages with both internal and external repositories. However, the engagement with external repositories fades after the round is raised. In column (4), we do not observe significant variations in the likelihood that a startup engages in a UI activity, pre- and post-round, both regarding own and external repositories. Finally, the results in column (5) show that startups do not increase their engagement with repositories related to best practices on HR management as they approach raising their first round. Overall, these findings suggest that startups – as they approach raising their first round – accelerate their investments into developing their technologies along the major components of the technology production process in order to attract funds by, especially, relying on external code repositories.

⟨ Insert Table 7 about here ⟩

One possibility is that startups engage with GitHub only to increase the visibility of their technologies before raising a round (Conti et al., 2013a,b; Hsu and Ziedonis, 2013) rather than to also develop and upgrade their innovation stack (as we have interpreted our findings so far). The fact that startups – prior to receiving a first round – intensify their activities on GitHub concerning all use-cases, but UI, serves as a first indication that increasing visibility is unlikely the sole driver of our findings. While SD/BE, ML, and API are fundamental features to ensure the functionality of a digital technology, UI serves as an out-ward facing building block. To provide further insight on the matter, we assess the dynamics of making new or previously private repositories public a the time startups raise a financing round. The slopes should be steeper before raising a round than after if the main purpose of engagement on GitHub is increasing visibility. In column 1 of Table 8, we examine the likelihood of making at least one new or previously private repository public, as measured by "Public Events" on the GitHub timeline of an organization account. As shown, we observe an increasing trend in making at least one repository public prior to raising a first financing round. Most

importantly, we do not detect any significant change in the trend after receiving financing (coefficient magnitude of $Post_{it} \times \tau_{it}$: -0.0002; p-value: 0.207). In column (2), we zoom in on the likelihood that a startup turns a previously private repository public and report a similar pattern as the one displayed in column (1). Because the dynamics of startups making their existing repositories public do not change before and after the first round, we infer that increasing the visibility of their technology to attract investors is unlikely the startups' sole motive for becoming involved with the GitHub open source community.[16]

Another possibility is that startups rely on the GitHub platform to add small tweaks to their existing technology in anticipation of a financing round, without substantial development. To assess this possibility, we consider the re-usage by other accounts of a startup's repositories. The results reported in column (3) of Table 8 show that the likelihood of publishing repositories that are eventually forked increases as the startup approaches its first financing round (coefficient magnitude of τ_{it}: 0.0002; p-value: 0.001), and its slope remains unchanged thereafter (coefficient magnitude of $Post_{it} \times \tau_{it}$: -0.0000; p-value: 0.352). As such, these findings suggest that the trajectory of publishing relevant repositories is similar before and after raising a first round.

⟨ Insert Table 8 about here ⟩

On the whole, these findings – together with our baseline results reported in Table 3 and in Figure 4 – indicate that startups are, indeed, "beefing up" their technology before they receive early-stage financing by relying on open source communities rather than merely increasing

[16]Note that this analysis does not rule out signaling, but suggests that signaling is unlikely the only explanation of the patterns we observe. As proposed in Spence (1973) and as examined by Conti et al. (2013a,b) in the context of entrepreneurship, signaling is a costly investment that entities make to convey specific information to uninformed parties. Because sending signals is costly, once startups have raised their financing round they should lower the intensity with which they invest in a signal, all else equal. In our empirical context, the signal would be making a repository public. There are positive costs associated with making repositories public, because, that way, startups reveal potentially valuable information regarding their technology to external parties. Given these costs and should GitHub be used for signaling purposes, startups should lower the intensity with which they make their repository public, once they have obtained funding. In Table 8, we observe exactly the opposite. The dynamics of a startup making private repositories public (column 2 of Table 8) are the same before and after raising a first financing round. This evidence, while clearly not conclusive, speaks against the possibility that signaling is the sole driver of our results.

their visibility to potential investors. Our results additionally indicate that – post-round – startups may become increasingly aware of appropriability issues, choosing less permissive open source licenses for their repositories.

4.2.2 Mechanisms

In this section, we explore the potential mechanisms driving our results. We begin by assessing whether a startup's engagement on GitHub varies depending on the type of technology a startup develops and its value proposition relative to competitors. We then evaluate whether such engagement is also detected prior to raising later-stage rounds. Finally, we examine heterogeneity in startup responses depending on the type of investors participating in a given round.

In Table 9, we investigate how the engagement on GitHub prior to raising a first financing round varies among startups that we classify as being software intensive according to the industry groups specified on Crunchbase. We omit from our models year by industry and year by region fixed effects as they may absorb important sectoral variation. The rationale for this analysis is that GitHub is primarily used by companies for whom software development is a core business activity and, thus, we may expect stronger effects for these companies. Consistent with this conjecture, the results in columns (1) and (2) show that "software startups" (column (2)) become relatively more involved on GitHub prior to raising funding than the other startups (column (1)), the difference among the coefficients of τ_{it} being statistically significant. After both sets of startups raise their first round, their increase in involvement on the GitHub platform decelerates.[17]

In columns (3) and (4) of Table 9, we assess how a startup's engagement on GitHub varies depending on the value proposition of the startup and its resulting differentiation relative to market incumbents. For this purpose, we employ the measure of strategic differentiation built by Guzman and Li (2021). To assess whether and how startups differentiate from

[17] Analyses reported in Table A8 show that the increased involvement of software-intensive startups with GitHub as they approach their first financing round is mostly driven by the use-cases of SD/BE, ML, and API. This suggests that the startups' activities we observe on GitHub may, indeed, closely map the technologies these startups develop.

market incumbents, the authors of the study apply natural language processing tools to the startups' online marketing statements. The authors then compute a differentiation score for each startup, calculated as the average distance between a startup's marketing statement and that of the startup's five closest competitors. The number of observations we employ for this analysis is much lower than for the other analyses as the the differentiation measure is available for a limited number of startups with accessible marketing statements. As shown, startups with a high level of differentiation with respect to incumbents (column (4)) become increasingly more involved on GitHub prior to raising a first round, relative to startups with a low differentiation level (column (3)). After startups raise their first round, we observe a leveling off of GitHub activities regardless of the positioning relative to competitors. However, these coefficients are non-significant on conventional levels provided the much smaller sample size. Taken together, these findings suggest that variations in a startup's value proposition might explain some of a startup's GitHub dynamics.

⟨ Insert Table 9 about here ⟩

The relevance of a startup's engagement on GitHub may vary depending on whether a startup seeks to raise a first financing round or subsequent rounds. While investors participating in a startup's first round may value a startup's technology investments relatively more, investors participating in follow-on rounds may prefer other aspects, such as a startup's marketing efforts (Wasserman, 2003). As a result, a startup's involvement with the GitHub community may matter more during early rounds and progressively fade as additional rounds are raised. To assess this conjecture, we re-estimate Eq. (2) for a startup's second and third round, respectively. The results are reported in Table 10 and in Figure 8. We adopt the same specification as the one reported in column (4) of Table 3. As shown, the increasing trend in the likelihood that a startup engages in a public activity on GitHub exhibits the steepest slope in the twelve months prior to raising a *first* round and is relatively flat in the twelve months prior to raising a *third* round. These findings are confirmed by analyses reported in Table A9, where we distinguish between Seed, Series A, and Series B to E rounds. Here,

we show that startups engage with the GitHub open source community particularly prior to raising a Seed round.

⟨ Insert Table 10 and Figure 8 about here ⟩

We next assess how a startup's engagement on GitHub varies with the type of investor participating in a startup's first round (Dushnitsky and Shaver, 2009). Specifically, we generate an indicator for whether at least one of the participating investors in a startup's first round is a VC. The reason for focusing on VCs is that they have been found to positively value a startup's technology relative to other investors (Conti et al., 2013b) and to be especially active in screening and nurturing their portfolio startups (Bernstein et al., 2016; Fitza et al., 2009; Sørensen, 2007). To investigate this potential source of heterogeneity, we modify Eq. (2) adding the interactions between VC and $Post_{it}$ and VC and τ_{it}, as well as the triple interaction between VC, $Post_{it}$, and τ_{it}. The results are reported in column (1) of Table 11.

As shown, the likelihood that a startup engages in a public activity on GitHub increases as the startup approaches the first round date, regardless of whether a VC participates in the round or not. However, the increment is significantly larger when a VC participates in the round. Post-round, the slope of the time trend declines for both VC-led and non-VC-led rounds.

A related aspect we consider is whether a startup's GitHub activities vary depending on how successful the investors participating in a startup's first financing round are. We examine two measures of investor success. The first is the number of investments an investor made in the five years prior to an observed startup's first round. The second is the number of successful investments made during the same time period. We define successful investments as the backing of a startup that ultimately exits via an acquisition or an IPO. For each observed round, we retained the maximum number of investments (or successful investments) made by the investors participating in the round. The results are reported in columns (2) and (3) of Table 11. As shown, the slope of the likelihood that a startup engages in a public

activity on GitHub prior to raising a first financing round is steeper the larger the number of past (successful) investments made by the most successful investor in a round.

⟨ Insert Table 11 about here ⟩

5 GitHub as a Tool for Strategy Research

Beyond the results that we report, and besides using GitHub as a way to store or access data and code used in research for replication purposes (Felten et al., 2021; Miric et al., 2022; Raffiee et al., 2022), our study shows that GitHub data has strong potential to help shed light on important questions in Strategy research.

One particularly salient example is to empirically assess theory derived from the resource based view (RBV) where firm resources can be defined as "all assets, capabilities, organizational processes, firm attributes, information, knowledge, etc. controlled by a firm that enable the firm to conceive of and implement strategies that improve its efficiency and effectiveness" (Barney, 1991). Established firms usually hold numerous resources of varying quality, whereas startups are oftentimes formed around few specific resources (Hsu and Ziedonis, 2013). Over time, the resource stock grows and resources are combined in a myriad of ways to support a firm's efforts – guided and constrained by market competition and competitor dynamics (Dierickx and Cool, 1989; Jacobides and Winter, 2012). Such flow of resources can be viewed to constitute a fundamental part of the strategy of the firm.

In a sense, GitHub data provides a unique window into the process of accumulating, recombining, redeploying, and discarding resources provided certain features of the data. Though GitHub data alone do not encompass the entire span of resources, they are particularly well-suited to capture those related to knowledge. Some of the features that enable this are: 1) the granularity of available information, 2) the ability to make temporal links, 3) relevance, and 4) transparency of use.

The granularity of information: GitHub provides extremely detailed information on the patterns and content of changes made to code. This can go as far as to a single word or digit.

The ability to make temporal links: Since activities on GitHub are time-stamped, researchers can study important dynamics associated with knowledge sourcing, development, redeployment and abandonment. Typically, researcher have been limited in their ability to see such precision when it comes to the timing of adjustments made to a firm's knowledge stock.

The relevance to the industry: As mentioned, GitHub is "the place where the world builds software" and the largest host of source code with over 40 million public repositories to date. Rarely can researchers get access to data that covers such a large number of a specific population in a single place. Moreover, those startups that use GitHub are highly relevant. As mentioned by Lin and Maruping (2022), 15 of the top 20 unicorns have a public GitHub repositories, averaging 62 public repositories each. These firms are Bytedance, Stripe, Didi, Chuxing, Instacart, Klarna, Epic Games, Databricks, Nubank, DJI Innovations, SHEIN, Checkout, Canva, Grab, Plaid, and BYJU's and are all digital startups that do not sell open source products or services.

The transparency of use: Given the requirements of open source, the data generating process on GitHub is transparent. Once a repository is public, everyone can see the changes being made to the code and observe the historic development of a repository.

6 Discussion and Conclusions

In this paper, we analyze unique data linking Crunchbase profiles to accounts on the software development platform GitHub. We investigate the role of open source technology investments among nascent firms in raising funds. Our results suggest that participation in open source communities plays an important role in achieving funding milestones. Specifically, we observe an acceleration in activities on GitHub as a firm approaches its first – especially, seed – financing round; an effect that decelerates in months thereafter. These results, which we obtain by controlling for fixed differences among startups and their technologies, life cycle effects, and technology shocks provide an indication that a startup's involvement with open source communities may be crucial for attracting earliest-round investors.

It is particularly noteworthy that startups intensify GitHub activities that relate to their own repositories less than GitHub activities related to external repositories. The most prevalent use-cases of external repositories pertain to software development/backend, machine learning and API. These findings suggests that startups may be using code repositories available on GitHub to scale and produce a minimal viable product necessary to attract funding.

Following an exogenous change in the Github price scheme which increased the cost of accessing external repositories versus internal collaboration, we show that startups intensify the level of internal collaborations in the months preceding their first round. This result may suggest that open source platforms offer relevant knowledge and technologies which startups access on the verge of attracting early-stage financing when the costs of doing so are comparatively low.

While startups appear to rely on open source communities to attract funding especially when relative costs are low, there are potential important trade-offs with regards to appropriability that startups make prior to raising the first round (Buss and Peukert, 2015; Teece, 1986). The trade-off between open-sourcing and appropriability is particularly relevant for technology startups (Gans and Stern, 2003). Because of their youth, small size, and resource constraints, startups are likely to gain by drawing from external sources of ideas. However, this reliance on external knowledge may make them vulnerable, especially to intellectual property concerns. Our finding that the likelihood of adopting permissive licenses levels off after a startup raises its first round suggests that with financing these concerns may become more central.

The results we present could also be interpreted such that startups use open source communities to increase their visibility. Startups may rely on these communities to signal the quality of their technology similar to individuals who contribute to online question-and-answer communities to capture recruiters' attention (Xu et al., 2020). It could then well be that signaling rather than technology production/improvement drives engagement on an open

source platform. Though a feasible interpretation, the evidence we provide seems to suggest that the increased involvement of startups on GitHub improves a startup's innovation pipeline and GitHub does not merely act as an amplifier of visibility. For one, startups are especially involved in activities that are crucial for the internal development and scalability of the technology. For another, the dynamics of making their existing private repositories public do not change before and after the first round, although publishing repositories should be the easiest and fastest way to show activity on GitHub. Moreover, the likelihood of publishing repositories that are eventually re-used by other account owners increases as a startup approaches its first financing round and its slope does not change thereafter, suggesting that the trajectory of publishing relevant repositories is similar before and after raising a first round. As such, and on the whole, our results seem to indicate that startups are, indeed, substantively "beefing up" their technologies before they receive early-stage financing by relying on open source communities.

Overall, this study contributes to increasing our understanding of the role of a particular channel through which outside knowledge can be accessed – open source – , what and how specific aspects of the components that are sourced matter for attracting funding, and investor preferences. As such, our findings extend the literature that analyzes firm commercialization strategies of open source software (Fosfuri et al., 2008), and the role of open source for firm productivity (Nagle, 2018b, 2019; Shah and Nagle, 2019). We highlight a novel channel through which startups benefit from using and actively engaging in open source software endeavors. Namely, in our context, technology startups rely heavily on external sources to develop – "beef up" – their own technologies and their involvement in open source communities matters for attracting investors, particularly VCs and successful investors, during startups' early stages. Further, we contribute to the entrepreneurial finance literature that has investigated whether VCs invest in the founding team or the technology (Bernstein et al., 2017; Gompers et al., 2020; Kaplan et al., 2009). We provide evidence that the focus of open source engagement lies in development, scale and integration, rather than in

the user interface, at least for raising the first round of financing. Finally, our use of machine learning algorithms to classify startups' activities on GitHub builds on an emerging line of research that applies sophisticated data techniques to categorize firm strategies (Conti et al., 2020; Guzman and Li, 2021).

The external validity of our approach may be limited given that we focus on a specific open source platform. However, GitHub is the largest host of source code with over 40 million public repositories to date and anecdotal evidence suggests that investors take public GitHub activities into consideration in their due diligence efforts (Jain, 2018). Although our results are based on particular activities on a specific online platform, we believe they have broader implications, especially for early stage ventures. Further, the fact that we can only observe public activities on GitHub implies that our results – especially those related to internal activities – are likely conservative estimates.

In conclusion, this paper provides important insight into entrepreneurial strategies for firms to attract financing. Our findings contribute to our understanding of the impact of early-stage tech-stack investments on achieving funding milestones (Roche et al., 2020). By opening the technology "black-box", we reveal important nuances that have been largely overlooked in the literature namely that using open source can substantially help firms attract funding, funds which are fundamental for the success of startups. Given the importance of entrepreneurship for economic growth (Adelino et al., 2017; Agarwal et al., 2007, 2010), these findings not only carry important implications for founders but also for policy-makers alike.

References

Adelino, M., Ma, S., and Robinson, D. (2017). "Firm age, investment opportunities, and job creation." *Journal of Finance*, *72*(3), 999–1038.

Agarwal, R., Audretsch, D., and Sarkar, M. (2007). "The process of creative construction: knowledge spillovers, entrepreneurship, and economic growth." *Strategic Entrepreneurship Journal*, *1*(3-4), 263–286.

Agarwal, R., Audretsch, D., and Sarkar, M. (2010). "Knowledge spillovers and strategic entrepreneurship." *Strategic entrepreneurship journal*, *4*(4), 271–283.

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

Barney, J. (1991). "Firm resources and sustained competitive advantage." *Journal of Management*, *17*(1), 99–120.

Bernstein, S., Giroud, X., and Townsend, R. R. (2016). "The impact of venture capital monitoring." *Journal of Finance*, *71*(4), 1591–1622.

Bernstein, S., Korteweg, A., and Laws, K. (2017). "Attracting early-stage investors: Evidence from a randomized field experiment." *Journal of Finance*, *72*(2), 509–538.

Bonaccorsi, A., Giannangeli, S., and Rossi, C. (2006). "Entry strategies under competing standards: Hybrid business models in the open source software industry." *Management Science*, *52*(7), 1085–1098.

Buss, P., and Peukert, C. (2015). "R&D outsourcing and intellectual property infringement." *Research Policy*, *44*(4), 977–989.

Choudhury, P., Allen, R. T., and Endres, M. G. (2021). "Machine learning for pattern discovery in management research." *Strategic Management Journal*, *42*(1), 30–57.

Conti, A., Guzman, J., and Rabi, R. (2020). "Information frictions in the market for startup acquisitions." *Available at SSRN 3678676*.

Conti, A., and Roche, M. P. (2021). "Lowering the bar? External conditions, opportunity costs, and high-tech start-up outcomes." *Organization Science*, *32*(4), 965–986.

Conti, A., Thursby, J., and Thursby, M. (2013a). "Patents as signals for startup financing." *Journal of Industrial Economics*, *61*(3), 592–622.

Conti, A., Thursby, M., and Rothaermel, F. T. (2013b). "Show me the right stuff: Signals for high-tech startups." *Journal of Economics & Management Strategy*, *22*(2), 341–364.

Dahlander, L. (2005). "Appropriation and appropriability in open source software." *International Journal of Innovation Management*, *9*(03), 259–285.

David, P. A., and Greenstein, S. (1990). "The economics of compatibility standards: An introduction to recent research." *Economics of Innovation and New Technology*, *1*(1-2), 3–41.

Dierickx, I., and Cool, K. (1989). "Asset stock accumulation and sustainability of competitive advantage." *Management Science*, *35*(12), 1504–1511.

Dushnitsky, G., and Matusik, S. F. (2019). "A fresh look at patterns and assumptions in the field of entrepreneurship: What can we learn?" *Strategic Entrepreneurship Journal*, *13*(4), 437–447.

Dushnitsky, G., and Shaver, J. M. (2009). "Limitations to interorganizational knowledge

acquisition: the paradox of corporate venture capital." *Strategic Management Journal*, *30*(10), 1045–1064.

Felten, E., Raj, M., and Seamans, R. (2021). "Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses." *Strategic Management Journal*, *42*(12), 2195–2217.

Fitza, M., Matusik, S. F., and Mosakowski, E. (2009). "Do VCs matter? the importance of owners on performance variance in start-up firms." *Strategic Management Journal*, *30*(4), 387–404.

Fosfuri, A., Giarratana, M. S., and Luzzi, A. (2008). "The penguin has entered the building: The commercialization of open source software products." *Organization Science*, *19*(2), 292–305.

Gans, J. S., and Stern, S. (2003). "The product market and the market for "ideas": Commercialization strategies for technology entrepreneurs." *Research Policy*, *32*(2), 333–350.

Germonprez, M., Kendall, J. E., Kendall, K. E., Mathiassen, L., Young, B., and Warner, B. (2017). "A theory of responsive design: A field study of corporate engagement with open source communities." *Information Systems Research*, *28*(1), 64–83.

Gompers, P. A., Gornall, W., Kaplan, S. N., and Strebulaev, I. A. (2020). "How do venture capitalists make decisions?" *Journal of Financial Economics*, *135*(1), 169–190.

Greenstein, S. (1996). "Invisible hands versus invisible advisors: Coordination mechanisms in economic networks." In E. Noam, and A. Nishuilleabhain (Eds.), *Public Networks, Public Objectives*, Amsterdam: Elsevier Science.

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

Greul, A., West, J., and Bock, S. (2018). "Open at birth? why new firms do (or don't) use open innovation." *Strategic Entrepreneurship Journal*, *12*(3), 392–420.

Guzman, J., and Li, A. (2021). "Measuring founding strategy." *Management Science, forthcoming*.

He, V. F., Puranam, P., Shrestha, Y. R., and von Krogh, G. (2020). "Resolving governance disputes in communities: A study of software license decisions." *Strategic Management Journal*, *41*(10), 1837–1868.

Hellmann, T., and Puri, M. (2002). "Venture capital and the professionalization of start-up firms: Empirical evidence." *Journal of Finance*, *57*(1), 169–197.

Hsu, D. H., and Ziedonis, R. H. (2013). "Resources as dual sources of advantage: Implications for valuing entrepreneurial-firm patents." *Strategic Management Journal*, *34*(7), 761–781.

Jacobides, M. G., and Winter, S. G. (2012). "Capabilities: Structure, agency, and evolution." *Organization Science*, *23*(5), 1365–1381.

Jain, V. (2018). "Investor due diligence: Beyond the obvious." `https://startupflux.com/investor-due-diligence-beyond-the-obvious/amp/`, accessed: 2021-6-29.

Jiang, J., Lo, D., He, J., Xia, X., Kochhar, P. S., and Zhang, L. (2017). "Why and how developers fork what from whom in github." *Empirical Software Engineering*, *22*(1), 547–578.

Kaplan, S. N., Sensoy, B. A., and Strömberg, P. (2009). "Should investors bet on the jockey or the horse? Evidence from the evolution of firms from early business plans to public companies." *Journal of Finance*, *64*(1), 75–115.

Lin, Y.-K., and Maruping, L. M. (2022). "Open source collaboration in digital entrepreneurship." *Organization Science*, *33*(1), 212–230.

Love, J. H., Roper, S., and Vahter, P. (2014). "Learning from openness: The dynamics of breadth in external innovation linkages." *Strategic Management Journal*, *35*(11), 1703–1716.

Miric, M., Jia, N., and Huang, K. G. (2022). "Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents." *Strategic Management Journal*.

Nagaraj, A. (2022). "The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry." *Management Science*, *68*(1), 564–582.

Nagle, F. (2018a). "Learning by contributing: Gaining competitive advantage through contribution to crowdsourced public goods." *Organization Science*, *29*(4), 569–587.

Nagle, F. (2018b). "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.

Raffiee, J., Fehder, D., and Teodoridis, F. (2022). "Revealing the revealed preferences of public firm ceos and top executives: A new database from credit card spending." *Strategic Management Journal*, *n/a*(n/a).

Roche, M., Oettl, A., and Catalini, C. (2020). "Entrepreneurs (co-) working in close proximity: Impacts on technology adoption and startup performance outcomes." *Harvard Business School Strategy Unit Working Paper*, (21-024).

Shah, S., and Nagle, F. (2019). "Why do user communities matter for strategy?" *Harvard Business School Strategy Unit Working Paper*, (19-126).

Shah, S. K. (2006). "Motivation, governance, and the viability of hybrid forms in open source software development." *Management Science*, *52*(7), 1000–1014.

Sheoran, J., Blincoe, K., Kalliamvakou, E., Damian, D., and Ell, J. (2014). "Understanding" watchers" on github." In *Proceedings of the 11th working conference on mining software repositories*, 336–339.

Sørensen, M. (2007). "How smart is smart money? A two-sided matching model of venture capital." *Journal of Finance*, *62*(6), 2725–2762.

Spence, M. (1973). "Job market signaling." *Quarterly Journal of Economics*, *87*(3), 355–374.

Stam, W. (2009). "When does community participation enhance the performance of open source software companies?" *Research Policy*, *38*(8), 1288–1299.

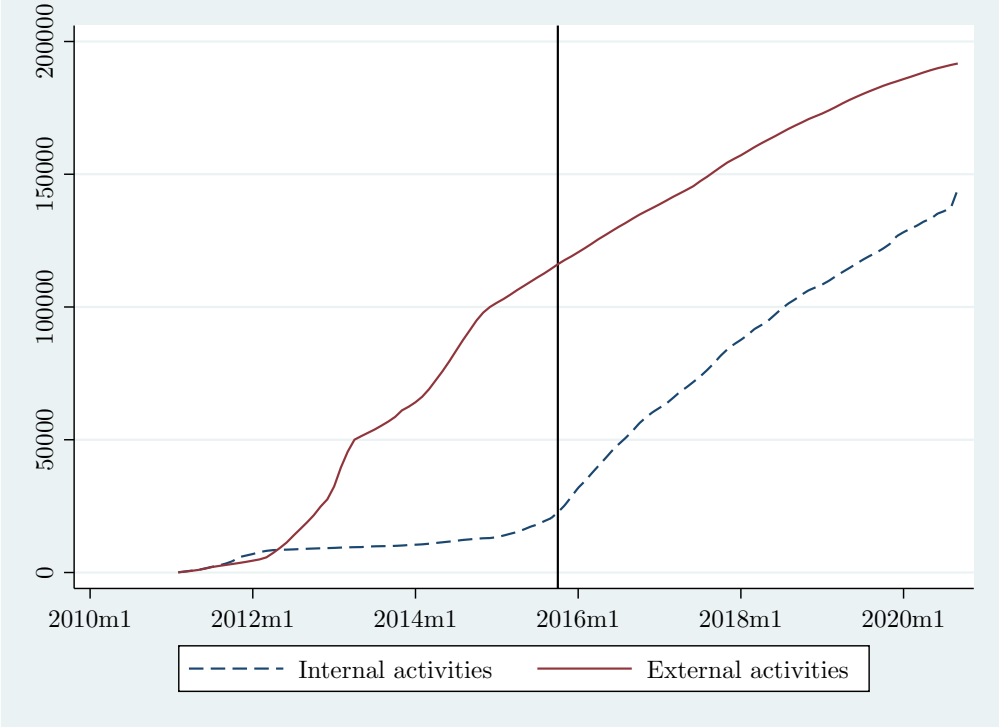
Teece, D. J. (1986). "Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy." *Research Policy*, *15*(6), 285–305.

Wasserman, N. (2003). "Founder-CEO succession and the paradox of entrepreneurial success." *Organization Science*, *14*(2), 149–172.

Xu, L., Nian, T., and Cabral, L. (2020). "What makes geeks tick? A study of stack overflow careers." *Management Science*, *66*(2), 587–604.

Yoo, Y., Boland Jr, R. J., Lyytinen, K., and Majchrzak, A. (2012). "Organizing for innovation in the digitized world." *Organization Science*, *23*(5), 1398–1408.

Figure 1: GitHub activities related to own and external repositories over time



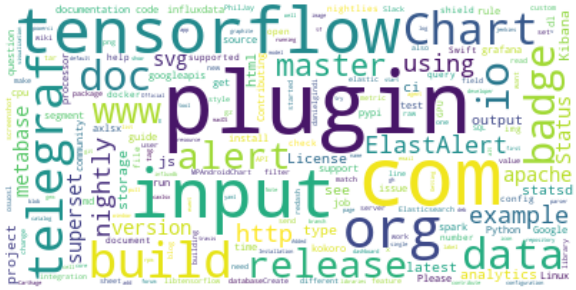
Notes: In this figure, we distinguish between a startup's public activities related to its own repositories (internal) and its engagement with external repositories (not controlled by the focal company). The vertical line indicates the time of the pricing change, which we exploit in a later section of the paper.

Figure 2: Repository classification: word clouds

Software (Back end), I

Software (Back end), II

Machine Learning



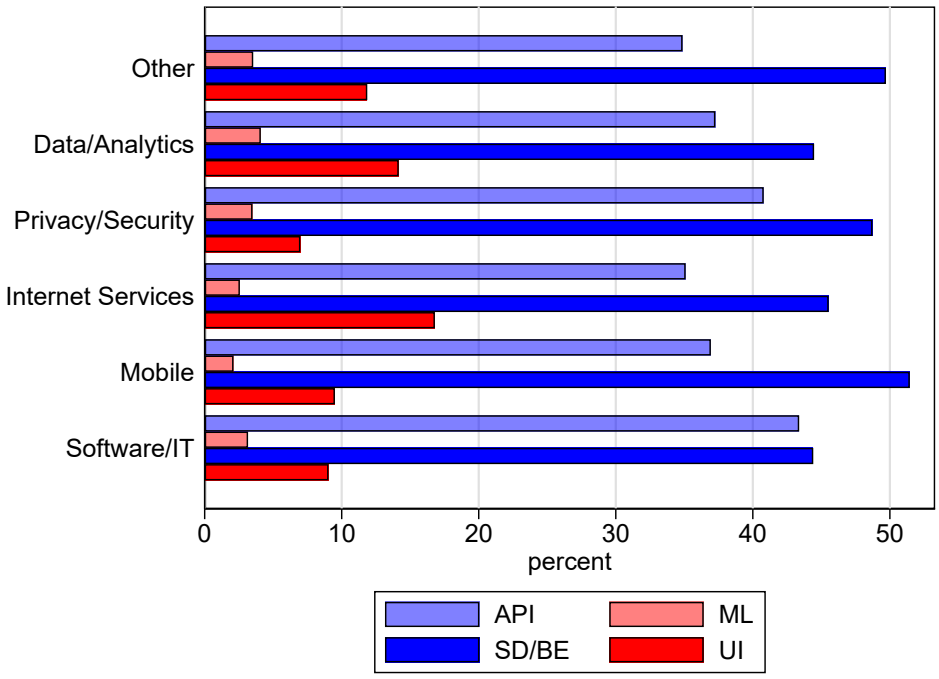
API

User Interface (UI)

Human Resources (HR)

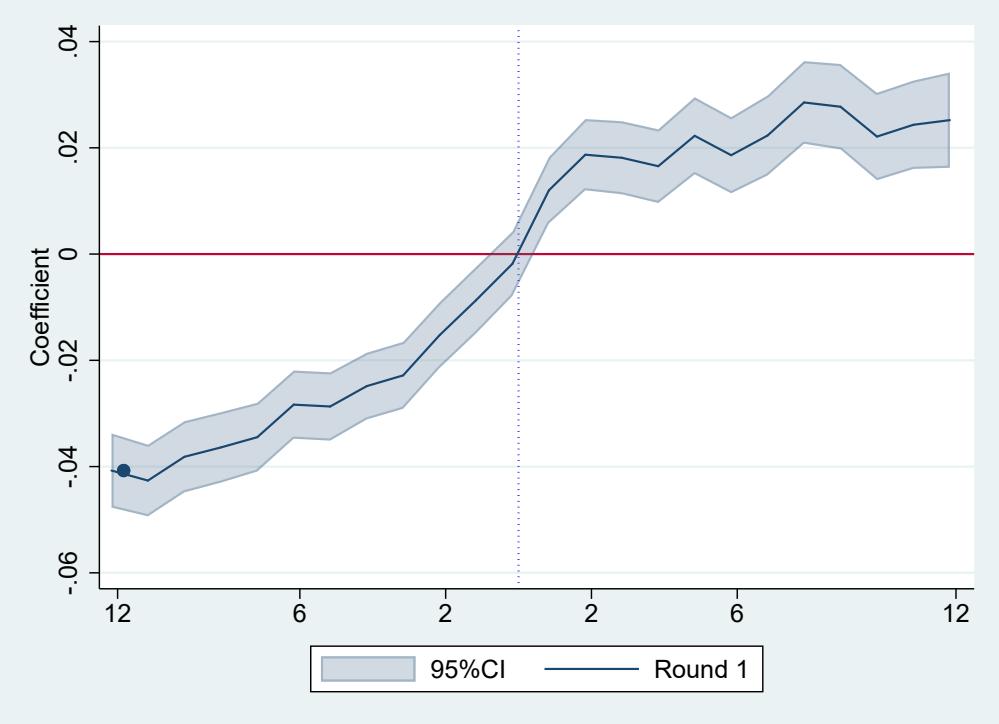
Notes: This figure displays the most common words for each of the use-cases we have considered. Note that these word clouds are based on the full classified sample, for which we used the hand-coded training sample as starting point for k-nearest neighbors. To construct this figure, word counts for a given category were discounted if the words also appeared frequently in other categories.

Figure 3: GitHub use-cases by industry groups



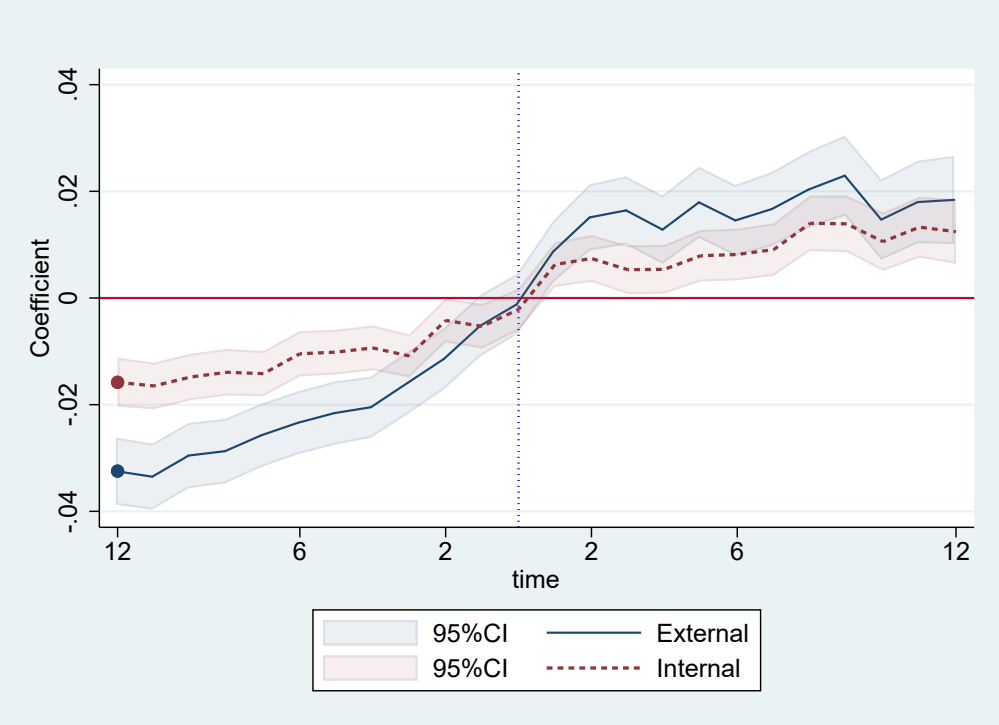
Notes: This figure reports the relevance of different use-cases across industry groups.

Figure 4: GitHub activity around the first round of financing



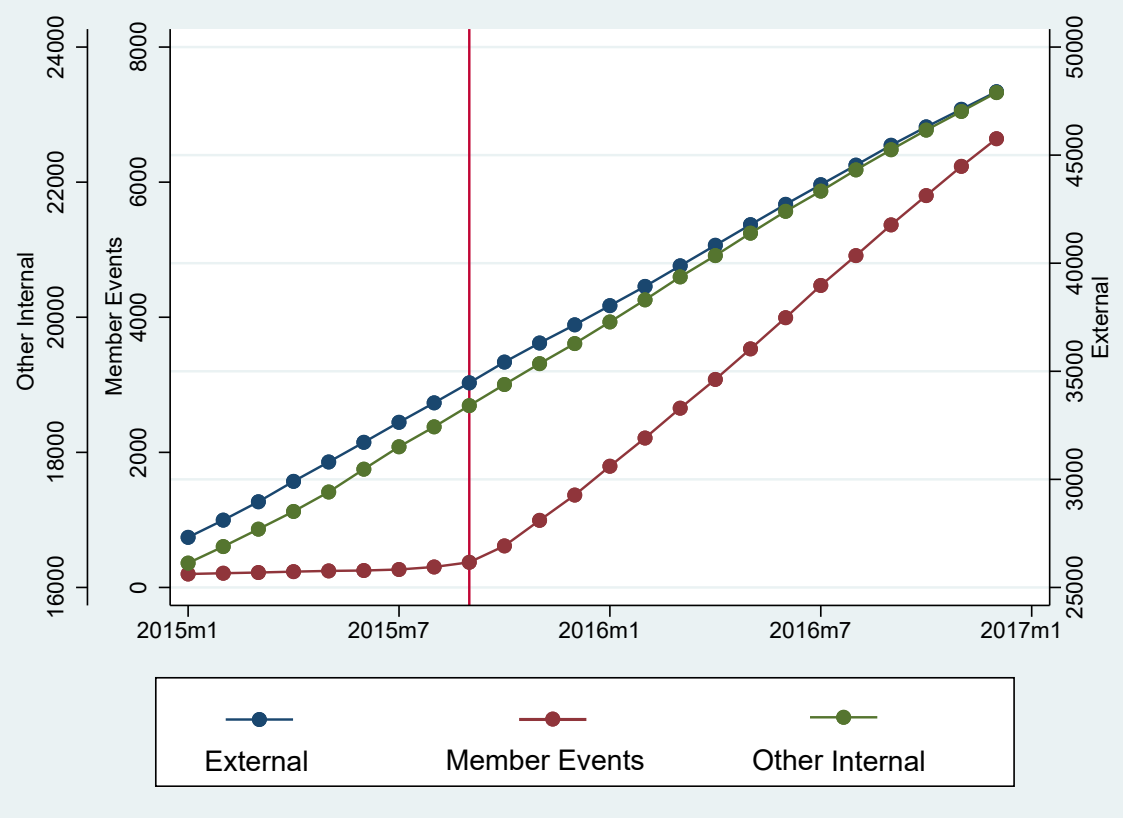
Notes: This figure displays how the engagement on GitHub prior to raising a first financing round and after varies among startups. The model is as specified in column (4) of Table 3. We replace τ with dummies for each month. The respective starting point coefficient is displayed in t-12 and the confidence interval is at the 95% level.

Figure 5: GitHub activity over time for external and internal activities



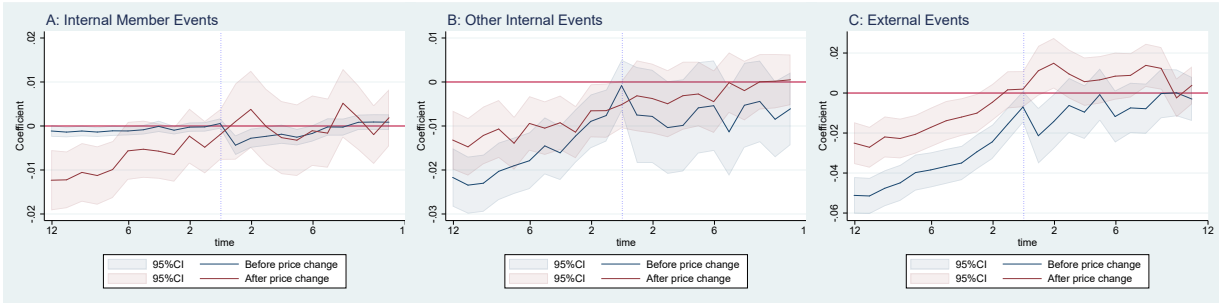
Notes: This figure displays how the engagement on GitHub prior to raising a first financing round and after varies among startups depending on whether the activities are external or internal. The solid line presents the results for all external activities and the dotted line displays results for internal activities. The models are as specified in column (4) of Table 3. We replace τ with dummies for each month. The respective starting point coefficient is displayed in t-12 and the confidence interval is at the 95% level.

Figure 6: GitHub activities at around the change of GitHub's pricing scheme



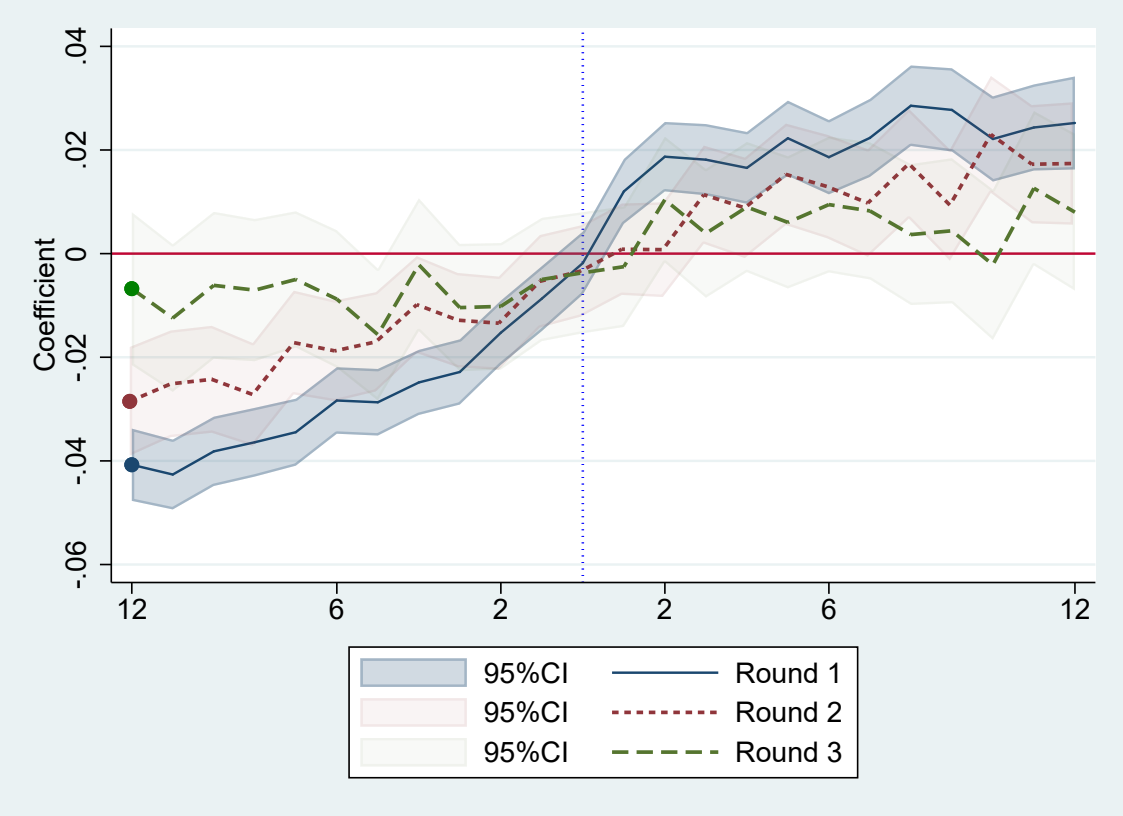
Notes: This figure displays the general changes of activity on GitHub depending on type at around the time of GitHub's change in the pricing scheme. In this graph we distinguish between external, internal without new member activity ("Other Internal") and internal new member activity only ("Member Events").

Figure 7: The effect of the change in GitHub's pricing scheme



Notes: This figure displays the effects of the change in GitHub's pricing scheme from a repository-based one to a user-based one. With the new pricing model, paying customers can create unlimited private repositories. We distinguish between internal member events (A), internal events without new member activity (B), and external events (C). The models are as specified in column (4) of Table 3. We replace τ with dummies for each month. The confidence interval is at the 95% level.

Figure 8: GitHub activity by round



Notes: This figure displays how engagement on GitHub prior to raising a first financing round and after varies among startups depending on the round in question. The underlying model is as specified in column (4) of Table 3. We replace τ with dummies for each month. The respective starting point coefficient is displayed in t-12 and the confidence interval is at the 95% level.

Table 1a: Summary statistics - Full sample of startups

	Obs	Mean	Std. Dev.	Min	Max	p50
Raised funds	160065	0.357	0.479	0	1	0
IPO	160065	0.012	0.107	0	1	0
Acquired	160065	0.081	0.273	0	1	0
GitHub	160065	0.093	0.29	0	1	0
Top Team	160065	0.002	0.039	0	1	0
AI	160065	0.039	0.194	0	1	0
Data Analytics	160065	0.098	0.297	0	1	0
Information Technology	160065	0.182	0.386	0	1	0
Internet Services	160065	0.200	0.400	0	1	0
Software	160065	0.356	0.479	0	1	0
N. Industry Groups	160065	3	2	1	19	3
Software share	154885	0.464	0.387	0	1	.5
California	160065	0.296	0.457	0	1	0
Massachusetts	160065	0.045	0.207	0	1	0
New York	160065	0.128	0.334	0	1	0

Table 1b: Top external and internal GitHub events

External			Internal		
Type	Total	Share	Type	Total	Share
ForkEvent	194518	0.96	MemberEvent	180338	0.46
WatchEvent	3486	0.02	PublicEvent	64355	0.17
PushEvent	1797	0.01	PushEvent	57196	0.15
Total Events (all types)	203565		Total Events (all types)	388001	

Table 1c: Summary statistics - Startups that raised funds and have a GitHub account

	Obs	Mean	Std. Dev.	Min	Max	p50
Information from GitHub:						
GitHub activity	10514	0.473	0.499	0	1	0
SD/BE	10514	0.227	0.419	0	1	0
ML	10514	0.042	0.201	0	1	0
API	10514	0.209	0.407	0	1	0
UI	10514	0.08	0.272	0	1	0
HR	10514	0.028	0.166	0	1	0
Information from Crunchbase:						
Artificial Intelligence	10514	0.109	0.312	0	1	0
Data Analytics	10514	0.227	0.419	0	1	0
Information Technology	10514	0.266	0.442	0	1	0
Internet Services	10514	0.285	0.451	0	1	0
Software	10514	0.615	0.487	0	1	1
N. Industry Groups	10514	3.711	1.803	1	15	3
Software share	10514	0.615	0.487	0	1	1
California	10514	0.434	0.496	0	1	0
Massachusetts	10514	0.059	0.236	0	1	0
New York	10514	0.159	0.366	0	1	0

Notes: The statistics reported regarding the various GitHub activities are in relation to the period starting twelve months prior to a startup's first financing round and ending twelve months after.

Table 2: Having a GitHub account: Determinants

	(1)	(2)	(3)	(4)
	Having a GitHub account			
Raised Funds$_i$	0.0779***	0.0769***	0.0870***	0.0442***
	(0.00510)	(0.00511)	(0.00545)	(0.00429)
Top Team$_i$		0.328***	0.337***	0.297***
		(0.0437)	(0.0440)	(0.0430)
Software share$_i$			0.108***	0.0775***
			(0.00452)	(0.00252)
Raised Funds$_i$ × Software share$_i$				0.0918***
				(0.00465)
Top Team$_i$× Software share$_i$				0.0657
				(0.107)
Founding year FE	Y	Y	Y	Y
Industry Group FE	Y	Y		
Observations	160065	160065	154885	154885
Mean DV	0.0930	0.0930	0.0944	0.0944

Notes: This table reports the results from estimating variants of Eq. (1). The dependent variable is the likelihood a startup had a GitHub organization account as of January 2021. The variable $RaisedFunds_i$ is an indicator that takes the value one if a startup raised at least one financing round and zero otherwise, while $TopTeam_i$ is an indicator identifying prominent founders and CXOs. The latter measure equals one if an employee is ranked among the top 1000 by Crunchbase and zero otherwise. We include fixed effects for a startup's founding year and for whether the startup is located in Massachusetts, New York, or California. In columns (1) and (2), we include industry group fixed effects. The industry groups we consider are Information Technology, Software, Data Analytics, Internet Services, and Artificial Intelligence. In columns (3) and (4), we replace our industry groups fixed effects with the share of a startup's industry group keywords that are related to software (Software share$_i$). The keywords related to software are: Apps, Artificial Intelligence, Consumer Electronics, Data and Analytics, Design, Financial Services, Gaming, Hardware, Information Technology, Internet Services, Messaging and Telecommunications, Mobile, Payments, Platforms, Privacy and Security, and Software. The number of observations in columns (3) and (4) is slightly lower than in columns (1) and (2) given that some startups do not have assigned industry keywords. Standard errors – reported in parentheses – are clustered at the startup founding year level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table 3: Startup engagement on GitHub

	(1)	(2)	(3)	(4)	(5)	(6)
	Engaging in a public activity on GitHub					
$Post_{it}$	0.0418***	0.0431***	0.0430***	0.0477***		
	(0.00549)	(0.00550)	(0.00550)	(0.00572)		
τ_{it}	0.00338***	0.00343***	0.00342***	0.00363***	0.00419***	0.00444***
	(0.000247)	(0.000248)	(0.000248)	(0.000249)	(0.000262)	(0.000365)
$Post_{it} \times \tau_{it}$	-0.00198***	-0.00209***	-0.00208***	-0.00253***	-0.00359***	-0.00321***
	(0.000377)	(0.000377)	(0.000377)	(0.000391)	(0.000487)	(0.000677)
Age_{it}	0.0745***	0.0675***	0.0661***			
	(0.00672)	(0.00673)	(0.00673)			
Startup FE	Y	Y	Y	Y	Y	Y
Year × Region FE		Y	Y	Y	Y	Y
Year × Industry Group FE		Y		Y	Y	Y
Year × Software share$_i$			Y			
Startup Age FE				Y		
Startup Age × Startup FE					Y	Y
$Post_{it}$ × Startup FE					Y	Y
Year × Lead Investor FE						Y
Observations	246797	246797	246797	246797	246797	152329
R2	0.282	0.283	0.283	0.284	0.429	0.429
Mean D.V.	0.082	0.082	0.082	0.082	0.082	0.090

Notes: This table reports the results from estimating variants of Eq. (2). $Post_{it}$ is equal to one for the twelve months that succeed a startup's first financing round, and zero in the twelve points preceding it. The variable τ_{it} is the count of months to/from a startup's first round. In column (1), we control for the natural logarithm of a startup's age. We additionally include startup and year fixed effects, as well as fixed effects controlling for whether a startup raised a second round in any of the twelve months following the first round. In column (2), we add region by year and industry group by year fixed effects. The industry groups we consider are Information Technology, Software, Data Analytics, Internet Services, and Artificial Intelligence. In column (3), we replace our industry groups with the share of a startup's industry group keywords that are related to software (*Software share*$_i$). The keywords related to software are: Apps, Artificial Intelligence, Consumer Electronics, Data and Analytics, Design, Financial Services, Gaming, Hardware, Information Technology, Internet Services, Messaging and Telecommunications, Mobile, Payments, Platforms, Privacy and Security, and Software. In column (4), we estimate a similar model as the one in column (3), this time replacing the natural logarithm of a startup's age with age fixed effects. We additionally interact each of the fixed effects for whether a startup raised a second round in any of the twelve months following the first round with the variable τ_{it}. In column (5), we estimate a similar model as the one in column (4) where we add age by startup and $Post_{it}$ by startup fixed effects. Finally, in column (6), we add lead investor by year fixed effects to the specification in column (5). Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table 4: Startup engagement on GitHub - Engagement with own repositories vs. engagement with external repositories

	(1)	(2)
	Engaging with:	
	Own repositories	External repositories
$Post_{it}$	0.0131***	0.0407***
	(0.00380)	(0.00523)
τ_{it}	0.00133***	0.00292***
	(0.000160)	(0.000226)
$Post_{it} \times \tau_{it}$	-0.000569**	-0.00225***
	(0.000255)	(0.000357)
Startup FE	Y	Y
Year × Region FE	Y	Y
Year × Industry Group FE	Y	Y
Startup Age FE	Y	Y
Observations	246797	246797
R2	0.261	0.254
Mean D.V.	0.033	0.066

Notes: This table reports the results from estimating variants of Eq. (2). The outcome in column (1) is whether a startup engages in public activities related to its own repositories in t. This measure proxies a startup's internal technology investments. The outcome in column (2) is whether a startup engages with external repositories in t. This outcome provides an indication of whether a startup builds on repositories controlled by other GitHub users to develop its technologies. $Post_{it}$ is equal to one for the twelve months that succeed a startup's first financing round, and zero in the twelve points preceding it. The variable τ_{it} is the count of months to/from a startup's first round. We include startup and startup age fixed effects, as well as fixed effects controlling for whether a startup raised a subsequent round in any of the twelve months following a first round. We further interact the latter fixed effects with τ_{it}. Additionally, we include region by year and industry group by year fixed effects. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table 5: Startup reaction to changes in GitHub's pricing scheme

	(1) Member event	(2) Other internal event	(3) External event
τ_{it}	0.000125***	0.00189***	0.00402***
	(0.0000456)	(0.000232)	(0.000315)
$Post_{it} \times \tau_{it}$	0.000201	-0.00134***	-0.00255***
	(0.000131)	(0.000468)	(0.000625)
$NewPriceScheme_t$	0.00507	0.00775**	0.00605
	(0.00346)	(0.00357)	(0.00581)
$Post_{it} \times NewPriceScheme_t$	0.0167**	-0.00647	0.0184
	(0.00806)	(0.00955)	(0.0143)
$NewPriceScheme_t \times \tau_{it}$	0.000902***	-0.000886***	-0.00140***
	(0.000232)	(0.000322)	(0.000481)
$Post_{it} \times NewPriceScheme_t \times \tau_{it}$	-0.000860**	0.000690	-0.000519
	(0.000394)	(0.000585)	(0.000830)
Startup FE	Y	Y	Y
Year × Region FE	Y	Y	Y
Year × Industry Group FE	Y	Y	Y
Startup Age × Startup FE	Y	Y	Y
$Post_{it}$ × Startup FE	Y	Y	Y
Observations	246797	246797	246797
R2	0.295	0.390	0.400

Notes: In this table, we assess how the introduction of a new GitHub pricing scheme in October 2015 affected startups' willingness to rely on open-source communities to develop their technologies and attract funds. We modify Eq. (2) including an indicator – $NewPriceScheme_t$ – identifying the period following the introduction of the new pricing scheme. We additionally introduce interaction terms between $NewPriceScheme_t$ and the following variables: $Post_{it}$, τ_{it}, and $Post_{it} \times \tau_{it}$. The dependent variables we consider are: an indicator that equals one if a startup engages in a member event in month t and zero otherwise (column 1); an indicator for whether a startup engages in a non-member, internal event (column 2); and an indicator for whether a startup engages with an external repository (column 3). We control for the full set of fixed effects. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table 6: Startup engagement on GitHub - New repositories with permissive licenses

	(1)	(2)	(3)	(4)	(5)	(6)
	New repositories with permissive licenses					
$Post_{it}$	0.0195***	0.0203***	0.0201***	0.0206***		
	(0.00402)	(0.00401)	(0.00402)	(0.00413)		
τ_{it}	0.00132***	0.00135***	0.00134***	0.00143***	0.00180***	0.00195***
	(0.000172)	(0.000172)	(0.000172)	(0.000173)	(0.000178)	(0.000241)
$Post_{it} \times \tau_{it}$	-0.000734***	-0.000794***	-0.000777***	-0.000902***	-0.00150***	-0.00155***
	(0.000270)	(0.000270)	(0.000270)	(0.000277)	(0.000346)	(0.000472)
Age_{it}	0.0474***	0.0439***	0.0438***			
	(0.00458)	(0.00456)	(0.00454)			
Startup FE	Y	Y	Y	Y	Y	Y
Year × Region FE		Y	Y	Y	Y	Y
Year × Industry Group FE		Y		Y	Y	Y
Year × Software share$_i$			Y			
Startup Age FE				Y		
Startup Age × Startup FE					Y	Y
$Post_{it}$ × Startup FE					Y	Y
Year × Lead Investor FE						Y
Observations	246797	246797	246797	246797	246797	152329
R2	0.182	0.183	0.182	0.183	0.329	0.326
Mean D.V.	0.039	0.039	0.039	0.039	0.039	0.042

Notes: This table reports the results from estimating variants of Eq. (2). The dependent variable is an indicator for whether a startup chooses a permissive license (e.g., BSD, MIT, Apache, CC-BY) for at least one new repository. $Post_{it}$ is equal to one for the twelve months that succeed a startup's first financing round, and zero in the twelve points preceding it. The variable τ_{it} represents the count of months to/from a startup's first round. In column (1), we control for the natural logarithm of a startup's age. We additionally include startup and year fixed effects, as well as fixed effects controlling for whether a startup raised a second round in any of the twelve months following the first round. In column (2), we add region by year and industry group by year fixed effects. The industry groups we consider are Information Technology, Software, Data Analytics, Internet Services, and Artificial Intelligence. In column (3), we replace our industry groups with the share of a startup's industry group keywords that are related to software (*Software share*$_i$). In column (4), we estimate a similar model as the one in column (3), this time replacing the natural logarithm of a startup's age with age fixed effects. We additionally interact each of the fixed effects with an indicator for whether a startup raised a second round in any of the twelve months following the first round with the variable τ_{it}. In column (5), we estimate a similar model as the one in column (4) where we add age by startup and $Post_{it}$ by startup fixed effects. Finally, in column (6), we add year by lead investor fixed effects to the specification in column (5). Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table 7: Startup engagement on GitHub - By GitHub activity

	Panel A: Own repositories				
	(1) SD/BE	(2) ML	(3) API	(4) UI	(5) HR
$Post_{it}$	0.00410**	0.00109*	0.00204	-0.000425	0.000168
	(0.00194)	(0.000569)	(0.00173)	(0.000880)	(0.000312)
τ_{it}	0.000369***	-0.0000145	0.000392***	0.0000104	-0.0000514***
	(0.0000776)	(0.0000213)	(0.0000685)	(0.0000287)	(0.0000163)
$Post_{it} \times \tau_{it}$	-0.000167	-0.0000591*	-0.000142	0.0000610	-0.00000236
	(0.000125)	(0.0000350)	(0.000115)	(0.0000545)	(0.0000204)
Observations	246797	246797	246797	246797	246797
R2	0.230	0.645	0.256	0.466	0.772
	Panel B: External repositories				
	(1) SD/BE	(2) ML	(3) API	(4) UI	(5) HR
$Post_{it}$	0.0110***	0.00184**	0.0102***	0.00235*	0.00163***
	(0.00263)	(0.000732)	(0.00249)	(0.00138)	(0.000519)
τ_{it}	0.000405***	0.0000762**	0.000316***	0.0000499	0.0000290
	(0.000114)	(0.0000345)	(0.000101)	(0.0000567)	(0.0000251)
$Post_{it} \times \tau_{it}$	-0.000599***	-0.000163***	-0.000584***	-0.0000749	-0.000126***
	(0.000170)	(0.0000489)	(0.000163)	(0.0000890)	(0.0000332)
Observations	246797	246797	246797	246797	246797
R2	0.189	0.464	0.185	0.280	0.599

Notes: This table reports the results from estimating variants of Eq. (2). The outcomes are whether a startup engages in public activities related to: software development/back end (SD/BE; column (1)); machine learning (ML; column (2)); Application Programming Interface (API; column (3)); user interface (UI; column (4)); human resources (HR; column (5)). In Panel A, we consider public activities on GitHub related to a startup's own repositories. In Panel B, we focus on a startup's engagement with external repositories. $Post_{it}$ is equal to one for the twelve months that succeed a startup's first financing round, and zero in the twelve points preceding it. The variable τ_{it} is the count of months to/from a startup's first round. We include startup and startup age fixed effects, as well as fixed effects controlling for whether a startup raised a subsequent round in any of the twelve months following a first round. We further interact the latter fixed effects with τ_{it}. Additionally, we include region by year and industry group by year fixed effects. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table 8: Making repositories public and relevance of an organization's repositories for external users

	(1)	(2)	(3)
	Making repositories public:		
	all	private	forked by others
$Post_{it}$	0.00649**	-0.000826	0.00270*
	(0.00281)	(0.00233)	(0.00155)
τ_{it}	0.000842***	0.000382***	0.000208***
	(0.000118)	(0.000130)	(0.0000610)
$Post_{it} \times \tau_{it}$	-0.000234	0.0000170	-0.0000947
	(0.000186)	(0.000171)	(0.000102)
Startup FE	Y	Y	Y
Year × Region FE	Y	Y	Y
Year × Industry Group FE	Y	Y	Y
Startup Age FE	Y	Y	Y
Observations	246797	244140	244140
R2	0.229	0.0865	0.130

Notes: This table reports the results from estimating variants of Eq. (2). In column (1), the dependent variable is an indicator for whether a startup makes a new or a previously private repository public. In column (2), the dependent variable is an indicator for whether at least one of the startups' repositories that were made public were created prior to the publication date. In column (3), the dependent variable is an indicator for whether at least one of the startups' repositories that were made public is forked by at least one other account. $Post_{it}$ is equal to one for the twelve months that succeed a startup's first financing round, and zero in the twelve points preceding it. The variable τ_{it} is the count of months to/from a startup's first round. In addition to the displayed fixed effects, we interact each of the fixed effects for whether a startup raised a second round in any of the twelve months following the first round with the variable τ_{it}. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table 9: Startup engagement on GitHub - Distinguishing startups by their technology and marketing statements

	(1) Software share$_i$ $\leq$Avg.	(2) Software share$_i$ >Avg.	(3) Differentiation$_i$ $\leq$Avg.	(4) Differentiation$_i$ >Avg.
$Post_{it}$	0.0240***	0.0464***	0.0278	0.0305*
	(0.00724)	(0.00786)	(0.0176)	(0.0163)
τ_{it}	0.00270***	0.00438***	0.00193**	0.00449***
	(0.000322)	(0.000371)	(0.000788)	(0.000750)
$Post_{it} \times \tau_{it}$	-0.00102**	-0.00232***	-0.000408	-0.00108
	(0.000489)	(0.000547)	(0.00120)	(0.00112)
Startup FE	Y	Y	Y	Y
Year FE	Y	Y	Y	Y
Startup Age FE	Y	Y	Y	Y
Observations	114938	131859	24168	27795
R2	0.246	0.301	0.306	0.269

Notes: This table reports the results from estimating Eq. (2). In column (1), we restrict the sample to companies whose value of the *Software share$_i$* variable is less than or equal to the mean. In column (2), we restrict the sample to companies whose value of the *Software share$_i$* variable is greater than the mean. In column (3), we restrict the sample to companies whose value of the *Differentiation$_i$* variable is less than or equal to the mean. In column (4), we restrict the sample to companies whose value of the *Differentiation$_i$* variable is greater than the mean. *Differentiation$_i$* is a measure of strategic differentiation. It is defined as the average distance between a startup's marketing statements and those of the five closest incumbents. The number of observations is lower in columns (3) and (4) than in columns (1) and (2) given that *Differentiation$_i$* is only available for a fraction of startups in the sample. The models include startup, age, and year fixed effects. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table 10: Startup engagement on GitHub - By round

	(1)	(2)	(3)
	Engaging in a public activity on GitHub		
	Round I	Round II	Round III
$Post_{it}$	0.0477***	0.0270***	0.0213**
	(0.00572)	(0.00462)	(0.00981)
τ_{it}	0.00363***	0.00321***	-0.000291
	(0.000249)	(0.000723)	(0.00272)
$Post_{it} \times \tau_{it}$	-0.00253***	-0.00131*	0.00160
	(0.000391)	(0.000722)	(0.00273)
Startup FE	Y	Y	Y
Year × Region FE	Y	Y	Y
Year × Industry Group FE	Y	Y	Y
Startup Age FE	Y	Y	Y
Observations	246797	169011	115296
R2	0.284	0.309	0.323

Notes: This table reports the results from estimating Eq. (2). $Post_{it}$ is equal to one for the twelve months that succeed a startup's given financing round, and zero in the twelve points preceding it. The variable τ_{it} is the count of months to/from a startup's round. In column (1), we examine a startup's engagement on GitHub before and after a first financing round. In column (2), we examine a startup's engagement on GitHub before and after a second financing round. Finally, in column (3), we examine a startup's engagement on GitHub before and after a third financing round. We include startup and startup age fixed effects, region by year and industry group by year fixed effects, as well as fixed effects controlling for whether a startup raised a subsequent round in any of the twelve months following the examined round. We interact these latter fixed effects with τ_{it}. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table 11: Startup engagement on GitHub - By investor type

	(1)	(2)	(3)
	Engaging in a public activity on GitHub		
	VC Round	Successful Investor Round	
$Post_{it}$	0.0466***	0.0428***	0.0425***
	(0.00687)	(0.00737)	(0.00694)
$Post_{it} \times \text{VC}_{it}$	0.000207		
	(0.0108)		
$Post_{it} \times \text{Success Inv.}_{it}$		0.000417***	
		(0.0000839)	
τ_{it}	0.00277***	0.00248***	0.00262***
	(0.000287)	(0.000305)	(0.000289)
$\tau_{it} \times \text{VC}_{it}$	0.00191***		
	(0.000419)		
$\tau_{it} \times \text{Success Inv.}_{it}$		0.000417***	0.000547***
		(0.0000839)	(0.000113)
$Post_{it} \times \tau_{it}$	-0.00243***	-0.00236***	-0.00245***
	(0.000464)	(0.000499)	(0.000471)
$Post_{it} \times \tau_{it} \times \text{VC}_{it}$	-0.0000497		
	(0.000725)		
$Post_{it} \times \tau_{it} \times \text{Success Inv.}_{it}$		0.000000282	0.0000237
		(0.000146)	(0.000194)
Startup FE	Y	Y	Y
Year × Region FE	Y	Y	Y
Year × Industry Group FE	Y	Y	Y
Startup Age FE	Y	Y	Y
Observations	246797	246797	246797
R2	0.285	0.285	0.285

Notes: This table reports the results from estimating variants of Eq. (2). $Post_{it}$ is equal to one for the twelve months that succeed a startup's first financing round, and zero in the twelve points preceding it. The variable τ_{it} represents the count of months to/from a startup's first round. We interact $Post_{it}$, τ_{it}, and $Post_{it} \times \tau_{it}$ with: an indicator that equals 1 if a startup had a VC participating in its first round (column (1)); the maximum number of investments in which any of a startup's investors participated in the five years preceding t (column (2)); the maximum number of successful investments made by a startup's investors in the five years preceding t (column (3)). A successful investment is one made into startups that ultimately exited via an acquisition or an IPO. We include startup and startup age fixed effects, region by year and industry group by year fixed effects, as well as fixed effects controlling for whether a startup raised a subsequent round in any of the twelve months following the examined round. We interact these latter fixed effects with τ_{it}. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Online Appendix for:

Beefing IT up for your Investor?
Open Sourcing and Startup Funding: Evidence from GitHub

Table A1: Having a GitHub account: Determinants

	(1)	(2)	(3)	(4)	5)
	Having a GitHub account				
Raised Funds$_i$	0.0769***	0.0760***	0.0751***	0.0743***	0.0735***
	(0.00511)	(0.00506)	(0.00499)	(0.00499)	(0.00496)
Top Team1000$_i$	0.328***				
	(0.0437)				
Top Team2000$_i$		0.308***			
		(0.0294)			
Top Team3000$_i$			0.292***		
			(0.0301)		
Top Team4000$_i$				0.280***	
				(0.0297)	
Top Team5000$_i$					0.278***
					(0.0272)
Founding year FE	Y	Y	Y	Y	Y
Industry Group FE	Y	Y	Y	Y	Y
Observations	160065	160065	160065	160065	160065

Notes: This table reports the results from estimating variants of Eq. (1). The dependent variable is the likelihood a startup had a GitHub organization account as of January 2021. The variable $RaisedFunds_i$ is an indicator that takes the value 1 if a startup raised at least one financing round and zero otherwise. $TopTeam1000_i$ is an indicator that equals one if an employee is ranked among the top 1000 by Crunchbase and zero otherwise. $TopTeam2000_i$ is an indicator that equals one if an employee is ranked among the top 2000 by Crunchbase and zero otherwise. $TopTeam3000_i$ is an indicator that equals one if an employee is ranked among the top 3000 by Crunchbase and zero otherwise. $TopTeam4000_i$ is an indicator that equals one if an employee is ranked among the top 4000 by Crunchbase and zero otherwise. $TopTeam5000_i$ is an indicator that equals one if an employee is ranked among the top 5000 by Crunchbase and zero otherwise. Standard errors – reported in parentheses – are clustered at the startup founding year level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table A2: Startup engagement on GitHub: Modifying the sample composition

	(1)	(2)	(3)	(4)	(5)	(6)
	Engaging in a public activity on GitHub					
$Post_{it}$	0.0497***	0.0494***	0.0496***	0.0545***		
	(0.00857)	(0.00858)	(0.00859)	(0.00892)		
τ_{it}	0.00134**	0.00131**	0.00133**	0.00153***	0.00198***	0.00151*
	(0.000533)	(0.000534)	(0.000534)	(0.000539)	(0.000640)	(0.000859)
$Post_{it} \times \tau_{it}$	-0.00272***	-0.00269***	-0.00273***	-0.00307***	-0.00411***	-0.00280**
	(0.000643)	(0.000644)	(0.000645)	(0.000664)	(0.000875)	(0.00118)
Age_{it}	0.0717***	0.0683***	0.0667***			
	(0.0158)	(0.0158)	(0.0159)			
Startup FE	Y	Y	Y	Y	Y	Y
Year × Region FE		Y	Y	Y	Y	Y
Year × Industry Group FE		Y		Y	Y	Y
Year × Software share$_i$			Y			
Startup Age FE				Y		
Startup Age × Startup FE					Y	Y
$Post_{it}$ × Startup FE					Y	Y
Year × Lead Investor FE						Y
Observations	135184	135184	135184	135184	135184	85596
R2	0.294	0.295	0.295	0.295	0.411	0.409

Notes: This table reports the results from estimating Eq. (2) having restricted the sample to startups that had a GitHub account at least twelve months prior to raising their first financing round. $Post_{it}$ is equal to one for the twelve months that succeed a startup's first financing round, and zero in the twelve points preceding it. The variable τ_{it} is the count of months to/from a startup's first round. In column (1), we control for the natural logarithm of a startup's age. We additionally include startup and year fixed effects, as well as fixed effects controlling for whether a startup raised a second round in any of the twelve months following the first round. In column (2), we add region by year and industry group by year fixed effects. The industry groups we consider are Information Technology, Software, Data Analytics, Internet Services, and Artificial Intelligence. In column (3), we replace our industry groups with the share of a startup's industry group keywords that are related to software (*Software share*$_i$). In column (4), we estimate a similar model as the one in column (3), this time replacing the natural logarithm of a startup's age with age fixed effects. We additionally interact each of the fixed effects for whether a startup raised a second round in any of the twelve months following the first round with the variable τ_{it}. In column (5), we estimate a similar model as the one in column (4) where we add age by startup and $Post_{it}$ by startup fixed effects. Finally, in column (6), we add year by lead investor fixed effects to the specification in column (5). Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table A3: Startup engagement on GitHub - Difference-in-differences model

	(1)	(2)
	Engaging in a public activity on GitHub	
τ_{it}	0.00133**	
	(0.000628)	
$\tau_{it} \times Post_{it}$	0.00000759	0.000178
	(0.00105)	(0.00105)
Has funds$_i \times \tau_{it}$	0.00274***	0.00266***
	(0.000716)	(0.000721)
Has funds$_i \times \tau_{it} \times Post_{it}$	-0.00325***	-0.00328***
	(0.00119)	(0.00119)
Startup FE	Y	Y
Treated-Control Group FE	Y	Y
Year × Region FE	Y	
Month × Region FE		Y
Year × Industry Group FE	Y	
Month × Industry Group FE		Y
Startup Age × Startup FE	Y	Y
$Post_{it}$ × Startup FE	Y	Y
R2	0.425	0.429

Notes: We estimate a difference-in-differences model comparing the dynamics of GitHub activities at around the time funded startups raise a first financing round to the GitHub dynamics of unfunded startups during the same time period. Control startups are randomly chosen from the set of startups that were founded during the same year and in the same state as the treated startups, and had a similar top team structure and share of software keywords. To each treated startup we assign up to ten controls. Standard errors – reported in parentheses – are clustered at the treated-control-group level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table A4: Startup engagement on GitHub: By activity type

	(1) Create event	(2) Fork event	(3) Gollum event	(4) Issue comment	(5) Member event	(6) Public event	(7) Pull request	(8) Push event	(9) Team event	(10) Watch event
$Post_{it}$	0.00699***	0.0343***	0.000364	0.000768	0.00411	0.00623**	0.000421	0.00107	0.00269*	0.000826
	(0.00182)	(0.00495)	(0.000650)	(0.000562)	(0.00277)	(0.00281)	(0.000569)	(0.00109)	(0.00157)	(0.000730)
τ_{it}	0.000218**	0.00255***	-0.0000543**	0.0000164	0.000868***	0.000882***	0.0000259	0.0000209	0.000306***	0.0000183
	(0.0000876)	(0.000211)	(0.0000216)	(0.0000287)	(0.000115)	(0.000118)	(0.0000260)	(0.0000469)	(0.0000614)	(0.0000330)
$Post_{it} \times \tau_{it}$	-0.000491***	-0.00184***	-0.00000568	-0.0000408	-0.000117	-0.000217	-0.0000299	-0.0000857	-0.000129	-0.0000720
	(0.000126)	(0.000337)	(0.0000385)	(0.0000382)	(0.000185)	(0.000185)	(0.0000403)	(0.0000800)	(0.000106)	(0.0000489)
Observations	246797	246797	246797	246797	246797	246797	246797	246797	246797	246797
R2	0.216	0.229	0.108	0.281	0.183	0.180	0.182	0.515	0.154	0.174

Notes: This table reports the results from estimating variants of Eq. (2) for the different types of a startup's engagement. In column (1), we examine the creation of a new event as an outcome. In column (2), we examine the forking of external repositories. In column (3), we analyze the likelihood that a startup creates a Gollum repository. In column (4), we analyze the issuing of a comment. In column (5), we analyze the creation of a member event. In column (6), we analyze the creation of a public event. In column (7), we analyze the creation of a pull request. In column (8), we analyze the creation of a push request. In column (9), we analyze the creation of a team event. In column (10), we analyze the creation of a watch event. A definition of the different events can be found at: https://docs.github.com/en/developers/webhooks-and-events/events/github-event-types (accessed March 2, 2022). We include the same fixed effects as those reported in column (4) of Table 3. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table A5: Engagements with external repositories: Share of total GitHub activities

	(1)	(2)	(3)	(4)	(5)	(6)
	Engagements with external repositories (%)					
$Post_{it}$	0.0204***	0.0211***	0.0209***	0.0234***		
	(0.00291)	(0.00291)	(0.00291)	(0.00299)		
τ_{it}	0.00149***	0.00152***	0.00151***	0.00161***	0.00190***	0.00200***
	(0.000127)	(0.000127)	(0.000127)	(0.000128)	(0.000134)	(0.000184)
$Post_{it} \times \tau_{it}$	-0.00106***	-0.00112***	-0.00110***	-0.00134***	-0.00190***	-0.00168***
	(0.000200)	(0.000200)	(0.000200)	(0.000206)	(0.000251)	(0.000348)
Age_{it}	0.0310***	0.0285***	0.0282***			
	(0.00344)	(0.00345)	(0.00346)			
Startup FE	Y	Y	Y	Y	Y	Y
Year × Region FE		Y	Y	Y	Y	Y
Year × Industry Group FE		Y		Y	Y	Y
Year × Software share$_i$			Y			
Startup Age FE				Y		
Startup Age × Startup FE					Y	Y
$Post_{it}$ × Startup FE					Y	Y
Year × Lead Investor FE						Y
Observations	246483	246483	246483	246483	246483	152123
R2	0.250	0.252	0.251	0.252	0.411	0.408

Notes: This table reports the results from estimating variants of Eq. (2). $Post_{it}$ is equal to one for the twelve months that succeed a startup's first financing round, and zero in the twelve points preceding it. The variable τ_{it} is the count of months to/from a startup's first round. In column (1), we control for the natural logarithm of a startup's age. We additionally include startup and year fixed effects, as well as fixed effects controlling for whether a startup raised a second round in any of the twelve months following the first round. In column (2), we add region by year and industry group by year fixed effects. The industry groups we consider are Information Technology, Software, Data Analytics, Internet Services, and Artificial Intelligence. In column (3), we replace our industry groups with the share of a startup's industry group keywords that are related to software (*Software share*$_i$). In column (4), we estimate a similar model as the one in column (3), this time replacing the natural logarithm of a startup's age with age fixed effects. We additionally interact each of the fixed effects for whether a startup raised a second round in any of the twelve months following the first round with the variable τ_{it}. In column (5), we estimate a similar model as the one in column (4) where we add age by startup and $Post_{it}$ by startup fixed effects. Finally, in column (6), we add year by lead investor fixed effects to the specification in column (5). Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table A6: Startup engagement on GitHub - All forks versus only forks of repositories with permissive licenses

	(1) All forks	(2) Forks of permissive repositories
$Post_{it}$	0.0343***	-0.000000432
	(0.00495)	(0.00114)
τ_{it}	0.00255***	0.0000315
	(0.000211)	(0.0000432)
$Post_{it} \times \tau_{it}$	-0.00184***	0.0000271
	(0.000337)	(0.0000718)
Observations	246761	246761
R2	0.229	0.0684

Notes: This table reports the results from estimating Eq. (2) for the likelihood that a startup forks an external repository (column (1)) and the likelihood that it forks an external repository with a permissive license (column (2)). We include startup, age, industry group by year and region by year fixed effects. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table A7: Startup reaction to changes in GitHub's pricing scheme

	(1) Member event	(2) External event	(3) Combined
τ	0.000125***	0.00402***	0.00405***
	(0.0000456)	(0.000315)	(0.000266)
$Post_{it} \times \tau_{it}$	0.000201	-0.00255***	-0.00166***
	(0.000131)	(0.000625)	(0.000527)
$NewPriceScheme_t$	0.00507	0.00605	0.00613
	(0.00346)	(0.00581)	(0.00416)
$Post_{it} \times NewPriceScheme_t$	0.0167**	0.0184	0.0182
	(0.00806)	(0.0143)	(0.0119)
$_{it} \times NewPriceScheme_t$	0.000902***	-0.00140***	-0.00116***
	(0.000232)	(0.000481)	(0.000412)
$Post_{it} \times \tau_{it} \times NewPriceScheme_t$	-0.000860**	-0.000519	-0.00120*
	(0.000394)	(0.000830)	(0.000724)
$MemberEvents_{it}$			-0.0158***
			(0.00168)
$Post_{it} \times MemberEvents_{it}$			-0.0273***
			(0.00772)
$\tau_{it} \times MemberEvents_{it}$			-0.00394***
			(0.000258)
$Post_{it} \times \tau_{it} \times MemberEvents_{it}$			0.000976*
			(0.000512)
$NewPriceScheme_t \times MemberEvents_{it}$			-0.00113
			(0.00250)
$Post_{it} \times NewPriceScheme_t \times MemberEvents_{it}$			-0.00126
			(0.00989)
$\tau_{it} \times NewPriceScheme_t \times MemberEvents_{it}$			0.00182***
			(0.000364)
$Post_{it} \times \tau_{it} \times NewPriceScheme_t \times MemberEvents_{it}$			0.00101
			(0.000663)
Observations	244856	244856	489712
R2	0.295	0.400	0.260

Notes: In this table, we assess how the introduction of a new GitHub pricing scheme in October 2015 affected startups' willingness to rely on open-source communities to develop their technologies and attract funds. The dependent variables we consider are: an indicator that equals one if a startup engages in a member event in month t and zero otherwise (column 1); an indicator for whether a startup engages with an external repository (column 2); an indicator for whether a startup engages with either a member event or an external repository (column 3). We control for the full set of fixed effects. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table A8: Startup engagement with external repositories: by software intensity

	External repositories				
	(1) SD/BE	(2) ML	(3) API	(4) UI	(5) HR
$Post_{it}$	0.00192	-0.00155	0.000822	-0.000285	0.00199*
	(0.00436)	(0.00117)	(0.00395)	(0.00239)	(0.00103)
τ_{it}	0.0000132	-0.0000130	-0.0000739	-0.0000746	-0.0000258
	(0.000178)	(0.0000506)	(0.000166)	(0.0000961)	(0.0000446)
$Post_{it} \times \tau_{it}$	0.0000150	0.0000597	-0.0000594	0.000138	-0.000108*
	(0.000281)	(0.0000782)	(0.000261)	(0.000151)	(0.0000638)
$\tau_{it} \times$ Software share$_i$	0.000641**	0.000146*	0.000640***	0.000204	0.0000900
	(0.000275)	(0.0000827)	(0.000248)	(0.000130)	(0.0000592)
$Post_{it} \times$ Software share$_i$	0.0148**	0.00557***	0.0154**	0.00434	-0.000574
	(0.00686)	(0.00209)	(0.00650)	(0.00355)	(0.00145)
$Post_{it} \times \tau_{it} \times$ Software share$_i$	-0.00101**	-0.000365***	-0.000864**	-0.000350	-0.0000299
	(0.000451)	(0.000140)	(0.000423)	(0.000225)	(0.0000941)
Observations	246797	246797	246797	246797	246797
R2	0.230	0.645	0.256	0.466	0.772

Notes: This table reports the results from estimating variants of Eq. (2). The outcomes are whether a startup engages in public activities related to: software development/back end (SD/BE; column (1)); machine learning (ML; column (2)); Application Programming Interface (API; column (3)); user interface (UI; column (4)); human resources (HR; column (5)). We focus on a startup's engagement with external repositories. $Post_{it}$ is equal to one for the twelve months that succeed a startup's first financing round, and zero in the twelve points preceding it. The variable τ_{it} is the count of months to/from a startup's first round. We include startup and startup age fixed effects, as well as fixed effects controlling for whether a startup raised a subsequent round in any of the twelve months following a first round. We further interact the latter fixed effects with τ_{it}. Additionally, we include region by year and industry group by year fixed effects. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

Table A9: Startup engagement on GitHub: Distinguishing between rounds

	(1) Seed	(2) Series A	(3) Series B to E
τ_{it}	0.00305***	0.00174**	-0.00127
	(0.000339)	(0.000845)	(0.00296)
$Post_{it}$	0.0456***	0.0281***	-0.00649
	(0.00467)	(0.00817)	(0.0236)
$Post_{it} \times \tau_{it}$	-0.00223***	-0.000406	0.00179
	(0.000364)	(0.000831)	(0.00296)
Startup FE	Y	Y	Y
Year × Region FE	Y	Y	Y
Year × Industry Group FE	Y	Y	Y
Startup Age FE	Y	Y	Y
Observations	218613	95462	103996
R2	0.256	0.310	0.327

Notes: This table reports the results from estimating Eq.(2). In column (1), we restrict the sample to Seed rounds. In column (2), we restrict the sample to Series A rounds. In column (3), we consider Series B to E rounds. Standard errors – reported in parentheses – are clustered at the startup level. Significance noted as: *p<0.10; **p<0.05; ***p<0.01.

A Repository classification

We classify the public repositories of all organizations, as well as the external repositories with which the organizations interacts through commits, pull requests or forks according to their type. We distinguish between repositories that pertain to software development/back end (SD/BE), machine learning (ML), application programming interface (API), and user interface (UI). To do so, we use the following methods:

A.1 TF-IDF vectorization

We first vectorize repository docs using the TF-IDF method. Each text document (that is, a collection of all the text in a single repository) is transformed into a vector of numbers. These numbers are the "scores" that the TF-IDF method assigns to each word in a document. A TF-IDF score is defined as the frequency with which a given word occurs in a certain document divided by the fraction of documents in which the word is present. More precisely, the formula used is: $tf * log(idf)$ where tf is the frequency of the word in a given document and idf is the inverse of the number of documents where the word appears. Thus, if a word is frequently used in a given document, it will have a high score. Similarly, if a word is used in few other documents it will also have a high score. Vectorization is fundamental to calculate the similarity between two repositories by comparing the vectors that represent them.

A.2 K-Nearest Neighbors prediction

Using the vectors produced by the TF-IDF method, we are able to compare any two repositories by calculating the dot product of the vectors that represent them. This calculation yields a similarity score between 0 and 1 (0 meaning totally different and 1 meaning totally similar).

Given that we can compare any two repositories, it is possible to classify any repository by examining similar repositories that have already been classified. To do so, we manually classify a subset of 150 repositories. We then use the KNN algorithm to classify the rest of the repositories using the 150 classified repositories as training data. The KNN algorithm consists of finding the k (in this case 5) most similar repositories in the training data and selecting the most common category to classify any new repository (each of the 5 similar repositories "votes" on a category for the new repository).

In order to accurately classify the large number of repositories we have available, we iteratively expand the training set applying the following steps:

1. Using the current training set, classify the rest of the repositories using KNN;
2. Retain only the repositories with the top N most confident classifications, and manually review them;
3. Add these repositories to the training set, and repeat the procedure until all repositories have been classified.

We increase the number N throughout the procedure as the training set grows, and with it, the accuracy of KNN. We calculate confidence using the number of "votes": if all 5 similar repositories "agree" on a category, then confidence is 100%. Conversely, if only 2 out of the 5 similar repositories "agree", then the confidence is low.[18]

[18] Due to the high number of categories, the "voting" process was weighted by distance, meaning votes from very similar documents counted more than votes from less similar documents.